How a teacher took on COVID-19 in India's Kerala state p. 752

Single-cell atlases of human development pp. 808 & 809

Fiber-optic sensors for multiple distortions p. 848

Science

$15
13 NOVEMBER 2020
SPECIAL ISSUE
sciencemag.org

AAAS

KEEPING COOL IN A WARMING WORLD

CONTENTS

13 NOVEMBER 2020 • VOLUME 370 • ISSUE 6518

756

Larvae of the emerald ash borer, an introduced beetle, have killed millions of ash trees in North America by tunneling beneath their bark.

NEWS

IN BRIEF

746 News at a glance

IN DEPTH

748 Effective vaccine offers shot of hope for pandemic
Pioneering RNA candidate seems to work well, although its makers revealed scant data *By J. Cohen*

749 A divided Congress could narrow Biden's ambitious plans
But president-elect has vowed to flex executive power *By J. Mervis*

750 To tackle pandemic, Biden must overcome distrust and division
President-elect vows "bedrock of science" will drive his administration's response to worsening U.S. crisis
By M. Wadman and W. Cornwall

751 New polio vaccine could boost faltering eradication drive
Sabin's oral vaccine has itself become a source of polio outbreaks. A modern, stabler version may reduce the risk *By L. Roberts*

752 India's COVID-19 teacher
Kerala health minister K. K. Shailaja has fought the pandemic with preparation, calm, and compassion *By V. Chandrashekhar*

754 Pandemic dooms Danish mink—and mink research
Mass culling ordered after discovery of mutations that may affect vaccine efficacy
By C. Lesté-Lasserre

755 Dust storms on Mars propel water's escape to space
Orbiters spy potential ongoing planetary "death spiral" *By P. Voosen*
RESEARCH ARTICLE p. 824

FEATURES

756 Rising from the ashes
An effort to breed North American ash trees that can resist a deadly beetle could accelerate the return of a forest icon *By G. Popkin*

INSIGHTS

POLICY FORUM

760 What's next for COVID-19 apps? Governance and oversight
Adaptive governance can help earn social license *By A. Blasimme and E. Vayena*

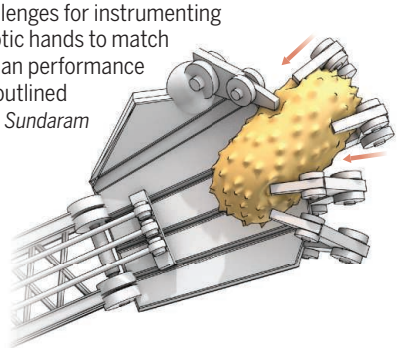
PERSPECTIVES

763 Understanding COVID-19 vaccine efficacy
Vaccine efficacy in high-risk groups and reduced viral shedding are important for protection *By M. Lipsitch and N. E. Dean*

765 Enhancing host cell infection by SARS-CoV-2
Neuropilin-1 binds the furin-processed spike protein of SARS-CoV-2 to promote virus entry *By M. Kielian*
REPORTS pp. 856 & 861

767 Peptide synthesis at the origin of life
Small-molecule organocatalysis might have driven the emergence of peptide biochemistry *By K. B. Muchowska and J. Moran*
REPORT p. 865

768 How to improve robotic touch
Challenges for instrumenting robotic hands to match human performance are outlined
By S. Sundaram



770 James S. Jackson (1944–2020)
Social psychologist who pioneered the study of Black culture *By T. A. Parham*

BOOKS *ET AL.*

771 Moving beyond the paradigm
A provocative 1975 call to leave behind dogmatic thinking could help scientists strengthen ties with the public *By J. Swift*

CREDITS: (PHOTO) MADDIE MCGARVEY; (GRAPHIC) C. BICKEL/*SCIENCE*

Published by AAAS

SPECIAL SECTION

Cooling technology

INTRODUCTION

776 Cooling in a warming world

NEWS

778 Living with heat *E. Pennisi*
PODCAST

PERSPECTIVES

782 Adapting to the challenges of warming
S. C. Sherwood

783 Cooling our insatiable demand for data *Amy S. Fleischer*

784 Photon-engineered radiative cooling textiles
P.-C. Hsu and X. Li
PODCAST

REVIEWS

786 Terrestrial radiative cooling: Using the cold universe as a renewable and sustainable energy source *X. Yin* et al.

791 New refrigerants and system configurations for vapor-compression refrigeration
M. O. McLinden et al.

797 Caloric materials for cooling and heating
X. Moya et al.

ON THE COVER

Air-conditioning units hang off the back side of a row of buildings in Shenyang, Liaoning province, China. Many Chinese people spend sanfu, the hottest period of the year, in air-conditioned buildings. As the planet continues to warm, the need for innovative and environmentally friendly cooling technologies becomes more important. See page 776.
Photo: Visual China Group via Getty Images

772 Imaginary demons and scientific discoveries
A new history acts as a field guide to the fictional imps invoked by scientific thinkers
By J. G. Michel

LETTERS

773 EcoHealth reframing of disease monitoring
By M. P. M. Vanhove et al.

773 Build international biorepository capacity
By J. P. Colella et al.

774 Response
By M. Watsa and Wildlife Disease Surveillance Focus Group

RESEARCH

IN BRIEF

805 From *Science* and other journals

RESEARCH ARTICLES

Human genomics

808 A human cell atlas of fetal gene expression *J. Cao* et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DX.DOI.ORG/10.1126/SCIENCE.ABA7721

809 A human cell atlas of fetal chromatin accessibility
S. Domcke et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DX.DOI.ORG/10.1126/SCIENCE.ABA7612

810 Plant science
Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions
J. R. Wendrich et al.
RESEARCH ARTICLE SUMMARY; FOR FULL TEXT: DX.DOI.ORG/10.1126/SCIENCE.AAY4970

811 Coronavirus
Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years *C. M. Saad-Roy* et al.

819 Plant science
Cell wall remodeling and vesicle trafficking mediate the root clock in *Arabidopsis*
G. Wachsman et al.

824 Martian atmosphere
Hydrogen escape from Mars is driven by seasonal and dust storm transport of water *S. W. Stone* et al.
NEWS STORY p. 755

REPORTS

831 Metallurgy
Constrained minimal-interface structures in polycrystalline copper with extremely fine grains *X. Y. Li* et al.

837 Cosmochemistry
Astronomical context of Solar System formation from molybdenum isotopes in meteorite inclusions *G. A. Brennecka* et al.

840 Optomechanics
Nano-acoustic resonator with ultralong phonon lifetime *G. S. MacCabe* et al.

844 Neuroscience
A thalamocortical top-down circuit for associative memory *M. B. Pardi* et al.

848 Sensors
Stretchable distributed fiber-optic sensors *H. Bai* et al.

853 Translation regulation
A phosphorylation-regulated eIF3d translation switch mediates cellular adaptation to metabolic stress
A. M. Lamper et al.

Coronavirus

856 Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity
L. Cantuti-Castelvetri et al.

861 Neuropilin-1 is a host factor for SARS-CoV-2 infection *J. L. Daly* et al.
PERSPECTIVE p. 765

865 Prebiotic chemistry
Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water *C. S. Foden* et al.
PERSPECTIVE p. 767

DEPARTMENTS

745 Editorial
The disease of distrust *By Reed V. Tuckson*

874 Working Life
Difficult conversations *By Sasha McDowell*

New Products.....870
Science Careers.....871

SCIENCE (ISSN 0036-8075) is published weekly on Friday, except last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue, NW, Washington, DC 20005. Periodicals mail postage (publication No. 484460) paid at Washington, DC, and additional mailing offices. Copyright © 2020 by the American Association for the Advancement of Science. The title SCIENCE is a registered trademark of the AAAS. Domestic individual membership, including subscription (12 months): $165 ($74 allocated to subscription). Domestic institutional subscription (51 issues): $2148; Foreign postage extra: Air assist delivery: $98. First class, airmail, student, and emeritus rates on request. Canadian rates with GST available upon request, GST #125488122. Publications Mail Agreement Number 1069624. **Printed in the U.S.A.**
Change of address: Allow 4 weeks, giving old and new addresses and 8-digit account number. **Postmaster:** Send change of address to AAAS, P.O. Box 96178, Washington, DC 20090–6178. **Single-copy sales:** $15 each plus shipping and handling available from backissues.sciencemag.org; bulk rate on request. **Authorization to reproduce** material for internal or personal use under circumstances not falling within the fair use provisions of the Copyright Act can be obtained through the Copyright Clearance Center (CCC), www.copyright.com. The identification code for *Science* is 0036-8075. *Science* is indexed in the *Reader's Guide to Periodical Literature* and in several specialized indexes.

Published by AAAS

EDITORIAL

The disease of distrust

An uncomfortable question has poked out from the chaos of the coronavirus disease 2019 (COVID-19) crisis—why does health inequity still persist in the United States? American people of color have been disproportionately affected by the pandemic. This, sadly, is no surprise because socioeconomic conditions, health care access challenges, and distrust in health care systems have historically prevented people of color from having healthier lives. As infections and deaths from COVID-19 continue to increase, effective treatments and vaccines are anxiously expected to become available soon. Unfortunately, less attention is being paid to questions about their equitable distribution and uptake. This only contributes to the suspicion felt by minority groups in the United States—particularly people of color—of the medical community. This barrier must come down. Every level of the health enterprise should pledge to reclaim the trust of all populations that is demanded by its professional oaths and missions.

This legacy of mistrust by people of color in health institutions, health professionals, researchers, and health policymakers in the United States has existed for decades. This painful reality has been amplified by the actions of the Trump administration and state-level officials, and by the behaviors of law enforcement, among others. It should be troubling to any health professional when thousands of African Americans pour into the streets to strenuously assert that their lives matter and that their humanity must be recognized. When people lose trust in the fundamental institutions of their society, decisions regarding the conduct of their lives become altered in both obvious and nuanced ways that affect their well-being.

I was privileged to serve as the Commissioner of Public Health for Washington, DC, during the height of the HIV-AIDS epidemic. This agency's work was substantially compromised by the legacy of the 40-year Tuskegee Syphilis Study, which began in 1932, where African American men were misled into participating in research to find a cure for the disease. This mistrust has been compounded by other incidents, including the failure of the biomedical community to acknowledge an African American woman, Henrietta Lacks, as the source of an invaluable research cell line. A statue across from the New York Academy of Medicine honoring James Marion Sims, who achieved success through research on un-anesthetized enslaved American persons, was another visible insult. Misperceptions regarding the etiology of disease, the motives of policy formulators, the veracity of health institutions, and the trustworthiness of the research enterprise have become almost as difficult to overcome as diseases themselves.

> "Misperceptions... have become almost as difficult to overcome as diseases themselves."

I am troubled by just how little the health profession has done to address the persistent misperceptions arising from the nation's history. Every aspect of the health enterprise must build the relationship between patients and health professionals. This includes facilitating inclusive input by disenfranchised communities into health policy formulation, reinforcing the actual and perceived protections and benefits of clinical research, and providing accessible scientific evidence to the public regarding therapeutics such as a COVID-19 vaccine.

Let this be the moment when the health community speaks to society in a manner that reassures the disenfranchised of the strength of the bond with them. When the human dignity of people is assaulted, health outcomes are affected, thereby requiring health professionals to speak out on matters of social concern. When patients receive care that deviates from best evidence because of bias or socioeconomic hurdles associated with structural racism, advocacy is required to recognize and address the problem. Whether through individual action or the collective work of professional societies, disenfranchised people need to be assured that in matters pertaining to their health, there can be confidence in scientific-based guidance and advice.

We must all recognize this disease of distrust as the scourge that it is and band together to reclaim this essential characteristic of the health profession: the preservation of the lives of all those who share our time and space. This should be the last time our society has to struggle against the legacy of the past as we fight persistent disparities in health outcomes and tackle this pandemic and the challenges to come.

–Reed V. Tuckson

PHOTO: REED TUCKSON

Reed V. Tuckson is the managing director of Tuckson Health Connections, LLC, a member of the U.S. National Academy of Medicine, and the former Commissioner of Public Health for the District of Columbia, USA. drreed@tucksonhealthconnections.com

10.1126/science.abf6109

Published by AAAS

NEWS

> “A mask remains the most potent weapon against the virus. Today’s news does not change that urgent reality.”
>
> **President-elect Joe Biden**, on 9 November following positive early results from a COVID-19 vaccine efficacy trial (see story, p. 748).

IN BRIEF

Edited by **Kelly Servick**

Hurricane Eta, which caused flooding in Central America last week, is among 29 named Atlantic Ocean storms this season.

METEOROLOGY

2020 brings stormiest Atlantic on record

With the rise of subtropical storm “Theta” this week, the Atlantic Ocean broke the record for the most named storms in a season, with 29—surpassing a high set in 2005. Unusually warm Atlantic waters and a La Niña event in the Pacific Ocean contributed to the stormy season. Forecasters are continually improving methods of identifying cyclones via satellite in the remote ocean, which makes the new record relevant only for the past few decades. Though it’s unclear whether global warming is leading to higher incidence of tropical cyclones, climate change is boosting rains and flooding left in the storms’ wake, and likely increasing their strength.

HIV shot proved for women

PUBLIC HEALTH | A long-lasting antiviral injection is 89% more effective at preventing HIV infections in women than a daily prophylactic pill, according to an announcement this week from the clinical trial network testing the approach. In a randomized, double-blind trial that included more than 3200 women across sub-Saharan Africa, scientists compared the approved pill marketed as Truvada—which many people struggle to take regularly—with an antiviral called cabotegravir, which was injected every 2 months. Whereas 34 participants taking Truvada became infected with HIV, just four receiving injections contracted the virus—a result that prompted the team to halt the trial early. Cabotegravir was previously shown to prevent HIV infections in cisgender men and transgender women who have sex with men.

Panel slams Alzheimer’s drug

DRUG DEVELOPMENT | An experimental Alzheimer’s disease treatment up for approval by the U.S. Food and Drug Administration (FDA) took a beating last week from an independent advisory committee convened to review its effectiveness. Aducanumab, an antibody drug developed by Biogen, clears away buildup in the brain of the protein beta amyloid, thought to drive neuronal damage. In 2019, Biogen halted two large clinical trials after it seemed neither would show that aducanumab slowed cognitive decline. Then, the company reversed course and sought FDA approval based on positive results in one of the trials. Of the advisory panel’s 11 voting members, 10 said the positive study did not provide enough evidence of effectiveness to support approval. (The 11th member voted

PHOTO: ORLANDO SIERRA/AFP VIA GETTY IMAGES

Published by AAAS

"uncertain.") FDA is expected to decide about aducanumab by March 2021; it doesn't have to follow the advice of its advisory committees, but it typically does.

Disputed replications fail again

PSYCHOLOGY | Ten high-profile psychology studies have failed to replicate for a second time, suggesting the failure of the first replication effort can't be explained by poor research methods, as its critics have argued. A 2015 study led by Brian Nosek at the Center for Open Science replicated only one-third of 100 seminal psychology findings. Nosek and colleagues have now led a second multilab effort to repeat 10 of the replication attempts that had not been endorsed by the original researchers. This time, the team had its research protocols peer reviewed—by those original researchers when possible—before launching its studies. Even though the revised protocols used study populations and methods closer to the original studies, they still didn't uphold the original findings. Those results were most likely unreliable, the team concludes in a paper published on 12 November in *Advances in Methods and Practices in Psychological Science*.

FDA must reassess GM salmon

BIOTECHNOLOGY | The U.S. Food and Drug Administration (FDA) must evaluate the hazard that genetically modified (GM) salmon pose to their wild relatives before it can approve any new rearing facilities, a U.S. district court ruled last week. In 2015, the fast-growing Atlantic salmon developed by AquaBounty—engineered to include genes from Pacific Chinook salmon and the eel-like ocean pout—became the first GM animal to be FDA approved for human consumption. Environmental groups sued FDA, claiming it had not adequately assessed the risks of the GM salmon escaping and interbreeding with wild salmon, especially endangered populations. U.S. District Judge Vince Chhabria found the agency had minimized risk by requiring that AquaBounty raise the salmon in land-based tanks. But he ruled that FDA must evaluate the possible consequence of escapes because approving more land-based facilities would increase escape risk. The decision does not affect current commercial production of the salmon, which began this summer.

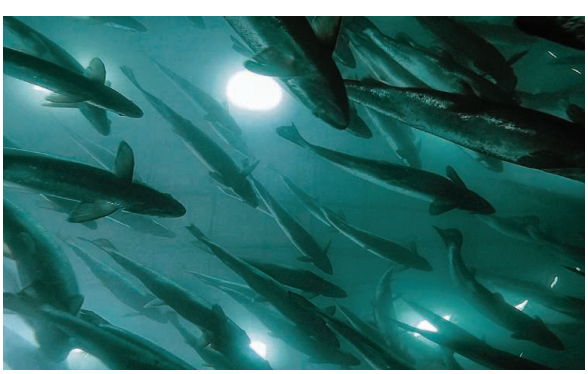


PHOTO: AQUABOUNTY

AquaBounty's fast-growing salmon, approved in 2015, faces a further regulatory hurdle before production facilities can expand.

Climate assessment turmoil

GEOSCIENCE | The Trump administration has abruptly reassigned Michael Kuperberg, head of the U.S. Global Change Research Program, which coordinates federal climate change research, including the National Climate Assessment (NCA). Kuperberg, who will return to the Department of Energy, will likely be replaced by David Legates, a climate contrarian the administration placed at the National Oceanic and Atmospheric Administration in September, *The New York Times* reported. The move came days after Elizabeth Weatherhead, a respected atmospheric scientist who accepts the reality of human-driven climate change, was named to lead the NCA, due in 2022. The incoming Biden administration can change these appointments, but the moves could delay the congressionally mandated NCA, used to guide climate policy.

New damage to Arecibo telescope

ASTRONOMY | The Arecibo Observatory was dealt a blow on 7 November, when one of the radio telescope's 12 main support cables snapped and damaged the 300-meter-wide main dish and other cables. The University of Central Florida, which leads a consortium managing the observatory on behalf of the National Science Foundation, is already working to repair damage from a previous cable break in August. Researchers now fear increased load on remaining cables could lead to a cascade of failures and the collapse of the antenna platform that is suspended over the dish. The nearly 60-year-old telescope, built into the hills of Puerto Rico, excels at studying pulsing radio sources in deep space and tracking nearby asteroids that could threaten Earth.

2020 ELECTION

Wins and losses for U.S. science

Most researchers hailed Joe Biden's victory in the race for U.S. president, but other election results delivered a mixed message about voter attitudes toward science.

Newbies founder

Democrat **Nancy Goroff's** bid to become the first female science Ph.D. in Congress appears doomed unless the Stony Brook University chemist, vying for a seat from New York in the U.S. House of Representatives, can win an overwhelming share of the votes not yet counted.

University of Wyoming marine ecologist **Merav Ben-David**, a Democrat, lost by almost a three-to-one margin in her race for an open U.S. Senate seat, and Democrat **Kathleen Williams**, a water resources expert, fell short in a second attempt to claim an at-large House seat from Montana.

Mixed fate for science figures

Two House members with science Ph.D.s, physicist **Bill Foster** (D–IL) and animal nutrition expert **Jim Baird** (R–IN), won by comfortable margins. Both serve on the House science committee, whose largely intact leadership hopes to build on a history of bipartisan legislation.

Most first-time House members with science and medical backgrounds kept their seats. But **Joe Cunningham** (D–SC), an ocean engineer, was ousted, as was **Kendra Horn** (D–OK), who chairs the science committee's space panel.

A narrow win by four-term Representative **Matt Cartwright** (D–PA) keeps him in the running to lead a House spending panel with jurisdiction over several science agencies.

Colorado senator **Cory Gardner**, the chief Republican sponsor of the 2016 American Innovation and Competitiveness Act that reauthorized a host of science programs, lost his seat to geologist and former Governor **John Hickenlooper** (D). Gardner's Democratic co-sponsor of the law, Senator **Gary Peters**, eked out a victory in Michigan.

State initiatives boost research

California voters appear to have narrowly approved $5.5 billion in stem cell research funding for the California Institute for Regenerative Medicine, pending completion of the count. Colorado voters gave the nod to planning the reintroduction of gray wolves, while Oregonians said yes to allowing researchers to test psilocybin, the active ingredient in so-called magic mushrooms, in treating a range of illnesses.

Published by AAAS

A dose of Pfizer and BioNTech's experimental COVID-19 vaccine is readied in October for a clinical trial participant in Turkey.

PHOTO: DOGUKAN KESKINKILIC/ANADOLU AGENCY VIA GETTY IMAGES

COVID-19

Effective vaccine offers shot of hope for pandemic

Pioneering RNA candidate seems to work well, although its makers revealed scant data

By **Jon Cohen**

On Sunday evening, standing beside the Hudson river in upstate New York, Kathrin Jansen was watching the sunset with her husband when the head of Pfizer, Albert Bourla, called with the update she was eagerly awaiting. Jansen, who heads vaccine R&D at Pfizer, had learned that morning a massive placebo-controlled trial of the company's candidate COVID-19 vaccine had a positive efficacy signal, but it wasn't clear just then how strong it was. Bourla told her an interim analysis found that the vaccine, which uses a potentially revolutionary RNA technology, was more than 90% effective at preventing symptomatic cases of COVID-19. "This is unbelievable," she told Bourla. Jansen and her husband began to jump up and down and soon had a champagne toast with "some really good stuff."

The rest of the world joined the celebration the next day as Pfizer and its German partner BioNTech announced the encouraging results of their candidate in a press release. Pfizer says it may seek an emergency use authorization (EUA) for the vaccine from the U.S. Food and Drug Administration (FDA) in a few weeks. But Peter Hotez, a vaccine researcher at Baylor College of Medicine, speaks for many scientists when he says that despite the "apparently good news," he would keep the champagne corked. "It's always hard to read the tea leaves of a company press release," without the underlying data, says Hotez, who is also part of a team making a vaccine against SARS-CoV-2, the virus that causes COVID-19. He stresses, too, that the vast majority of the public will not have access to this or any other COVID-19 vaccine for several months.

"There are a lot of unanswered questions," adds Georgetown University's Jesse Goodman, former chief scientist at FDA. Nothing is known about how long the immunity triggered by the vaccine will last, whether it can prevent severe COVID-19, and even whether it will slow transmission rates if it's used widely in a population. It's unclear how well it works in the elderly, who suffer the most from SARS-CoV-2. The vaccine, based on a simple strand of messenger RNA (mRNA), has to be stored at −80°C to preserve the genetic material, and making and delivering enough for hundreds of millions—if not billions—of people remain huge challenges.

BioNTech, a small company working on mRNA vaccines to treat cancer, designed the COVID-19 vaccine and then partnered with Pfizer. Ugur Sahin, BioNTech's CEO, says they tested more than 20 different stretches of mRNA coding for portions of spike, the surface protein the coronavirus uses to bind to human cells. They looked for mRNAs that were most likely to be taken up and expressed by dendritic cells, which "present" spike to the immune system to initiate antibody and T cell responses against the virus. "We spent a lot of effort in understanding what you need to get the vaccine into dendritic cells," Sahin says. "That's our key differentiator." The BioNTech team then chose the mRNA that triggered the best immune responses in early human trials.

The subsequent efficacy trial, which has enrolled more than 44,000 people, quickly yielded results because of the skyrocketing number of COVID-19 cases in the United States, home to the vast majority of the study's 154 sites. The design called for an independent panel to unblind the data if 62 participants had symptoms of COVID-19 and tested positive for SARS-CoV-2. The trial hit the 62 case mark on 3 November, Jansen says, and the companies started to carefully comb through the cases to prepare the data for the panel to evaluate Sunday morning, by which time the trial had 94 people with COVID-19.

The company has not revealed how many of those cases were in the vaccine group versus the placebo group. So far, the trial has found no major side effects. For safety reasons, FDA guidelines say before a company can seek an EUA, 2 months must pass after at least half of the participants in a vaccine trial have received all of their doses. That point is about 2 weeks from now, and Pfizer anticipates the trial will by then have reached its planned endpoint of at least 164 accumulated COVID-19 cases.

If Pfizer then files for an EUA, FDA has promised to hold an open meeting of its vaccine advisory board—a step committee members say is critical. "We all have

Published by AAAS

a healthy skepticism because companies clearly want to sell their vaccines," says Cody Meissner, a pediatrician at Tufts University School of Medicine who is on the committee and was nonetheless heartened by the positive report.

If the vaccine is given an EUA by FDA and regulators in other countries, or even formally approved, an ethical dilemma comes to the fore: whether to offer it to the trial participants who received the placebo, which could compromise collecting more data on the durability of the vaccine's immune response and safety. Participants in trials of other candidates might also drop out if they could access a 90% effective vaccine. "It could have a huge impact," says bioethicist Christine Grady of the U.S. National Institutes of Health, noting that fighting the pandemic may require multiple vaccines, with known strengths and weaknesses. "Unless we have good data, that's going to be a hard question to answer."

Pfizer and BioNTech developed their vaccine without help from Operation Warp Speed, a massive U.S. effort to fast-track COVID-19 vaccines, although the companies did agree to sell the government 100 million doses. Pfizer has cut deals with several other countries as well. Each person requires two doses, and Pfizer and BioNTech have produced only 500,000 doses, although they expect to have 50 million by the end of the year. By 2021, the companies project they can produce 1.3 billion doses.

Still, initial supplies will be severely limited even in the United States and they will remain short for most of the world next year. "If I look at the dose projections for 2021, and I look at what's now been spoken for in bilateral deals around the world, it creates a fair amount of anxiety," says Nicole Lurie of the Coalition for Epidemic Preparedness Innovations, a nonprofit that aims to improve equity and access to COVID-19 vaccines and drugs for all countries.

Seth Berkley, who heads Gavi, the Vaccine Alliance, says the positive vaccine data bode well for the many other candidates in trials or development that also rely on the SARS-CoV-2 spike protein, some of which don't need ultracold storage. It is also an encouraging sign for other vaccine candidates using mRNA—one of which, from Moderna, could also have efficacy data this month.

Lawrence Corey, a vaccine researcher at the Fred Hutchinson Cancer Research Center who heads a U.S. network that's testing several COVID-19 vaccines (but not the Pfizer/BioNTech candidate), calls it an "amazing feat" that a vaccine has a clear efficacy signal just 11 months after SARS-CoV-2 was identified. "This is a revelation of science," Corey says. ■

2020 U.S. ELECTION

A divided Congress could narrow Biden's ambitious plans

But president-elect has vowed to flex executive power

By **Jeffrey Mervis**

Most U.S. researchers and environmental activists were ecstatic when Joe Biden emerged as the winner of the U.S. presidential election on Saturday. They expect him to reverse a host of Trump administration policies they oppose and push for new steps to fight climate change.

But when Biden is sworn in on 20 January 2021, his ability to advance an ambitious agenda will be constrained by his likely status as the first president in more than 30 years to take office without his party controlling both chambers of Congress. Republicans are favored to preserve their majority in the Senate by winning at least one of the two runoff contests in Georgia, and Democrats will have a narrower majority in the new House of Representatives than during the previous 2 years.

New appointees should end the "denigration of expertise" under the Trump administration.

David Hart, George Mason University

Biden won't need a Democratic Congress to undo an array of executive orders issued by President Donald Trump over the past 4 years. The president-elect has said he will rejoin the Paris climate pact on his first day in office and cancel orders that weakened environmental regulations and barred immigrants from many majority-Muslim nations. His new administration will also likely suspend work on proposed regulations it opposes, including several weakening pollution controls. The halt will effectively kill them, but it could take years to reverse regulations that are already finalized.

Biden will also have substantial opportunity to reverse the "denigration of expertise" that permeated the Trump presidency, says science policy specialist David Hart of George Mason University. The outgoing administration's disregard for evidence-based policy resulted in attempts to undermine the accuracy of this year's census, politicize climate and hurricane forecasts, and sideline scientific advisory panels. It also triggered a "crisis of confidence" at federal regulatory agencies, says Daniel Sarewitz of Arizona State University, Tempe.

To restore that confidence, Biden will likely populate his administration with well-respected researchers. Most science agencies will be getting new leaders, although Sethuraman Panchanathan is only 5 months into a 6-year term as director of the National Science Foundation, and Francis Collins may be asked to extend his 11 years at the helm of the National Institutes of Health until the coronavirus pandemic is controlled.

Biden's choice of a science adviser and head of the White House Office of Science and Technology Policy could send a key signal about the status of science in his administration. The scope of that job, now held by Kelvin Droegemeier, was greatly diminished under Trump.

A Republican majority in the Senate would pose a serious obstacle to Biden's campaign promises aimed at combating climate change, including a $2 trillion green infrastructure initiative. Budget hawks in both parties are also expected to put pressure on their leaders to rein in spending after pandemic relief packages caused the federal deficit to skyrocket. Historically, however, Congress has backed increases in federal research spending even during periods of austerity.

A first look at Biden's priorities for science will come when the administration rolls out its 2021 budget request in February. Even before Biden is sworn in, however, leaders of the House and Senate have said they want to pass another pandemic relief bill, which could include tens of billions of dollars to help universities recover from the losses caused by shutting down campus labs in the spring.

Lawmakers must also decide whether to complete work on this year's annual budget. In July, the House approved bills containing healthy boosts for several science agencies, and this week the Senate released its preliminary numbers. If the two bodies can't come to an agreement, the alternative is to continue the current freeze on federal spending and leave the dealmaking to the new president and Congress. ■

Published by AAAS

In October, then-candidate Joe Biden received a COVID-19 briefing from public health experts. He recently named many to a new advisory board.

PHOTO: DREW ANGERER/GETTY IMAGES

2020 U.S. ELECTION

To tackle pandemic, Biden must overcome distrust and division

President-elect vows "bedrock of science" will drive his administration's response to worsening U.S. crisis

By **Meredith Wadman** *and* **Warren Cornwall**

President-elect Joe Biden wasted little time this week in following through on his promise to put science first in confronting the coronavirus pandemic, naming a 13-member COVID-19 advisory board of prominent researchers and public health experts and unveiling new details of his evidence-based battle plan.

Many observers, however, say executing that plan will require not just technical know-how, but also plenty of public persuasion and political savvy. And they worry the pandemic will get much worse before Biden takes office on 20 January 2021, further complicating efforts to end the crisis, which was growing by more than 100,000 cases a day and had killed at least 238,000 people in the United States as *Science* went to press.

"Biden's got a bigger challenge than any other country in the world," says Nahid Bhadelia, medical director of the Special Pathogens Unit at Boston Medical Center.

Biden said his advisory board, unveiled on 9 November, will rely on a "bedrock of science" and "help shape my approach to managing the surge in reported infections; ensuring vaccines are safe, effective, and distributed efficiently, equitably, and free; and protecting at-risk populations." It has three co-chairs: Marcella Nunez-Smith, an internist and health equity researcher at the Yale University School of Medicine; former Obama administration Surgeon General Vivek Murthy; and David Kessler, who ran the Food and Drug Administration under former Presidents George H. W. Bush and Bill Clinton.

The same day, Biden's transition team updated its existing pandemic plan. It now says Biden will push every governor to require people to wear face masks around anyone outside their household and will work to provide an additional $25 billion in federal funding to manufacture, distribute, and administer a free vaccine or vaccines. That pledge came as Pfizer announced that interim data from a key clinical trial showed its vaccine is likely effective (see p. 748).

The moves came just 48 hours after news networks declared Biden the winner of the election and highlighted his emphasis on fighting the pandemic. They were designed to show he will mount a far more muscular federal response than President Donald Trump, who has largely left matters to state and local officials. Among other things, Biden's plan calls for standing up a 100,000-person corps to trace people who have been in contact with an infected individual, massively increasing the availability of COVID-19 tests, and using the full power of the Defense Production Act to replenish depleted stocks of personal protective equipment.

Once sworn in, Biden will be able to quickly realize some of these plans through executive action. Other efforts, however, will require Congress to approve new funding or pass legislation. And winning such approval could take time, and prove difficult, if Republicans maintain control of the Senate. That will be decided by the results of two runoff elections in Georgia set for early January 2021.

In the meantime, observers hope Biden's team will begin to correct one of the Trump administration's biggest missteps: its chaotic, contradictory, and often false messaging, which has sown confusion about the course of the pandemic, bolstered opposition to masks, and raised fears that vaccines would be rushed through the approval process. Pew Research Center polling, for example, has found the number of Americans saying they would be willing to take a vaccine fell from 72% in May to 51% in September.

"The overarching issue is the need to rebuild trust" in what the federal government is saying, says Tom Frieden, CEO of Resolve to Save Lives, who directed the Centers for Disease Control and Prevention (CDC) under former President Barack Obama. "Social cohesion is essential to the control of infectious diseases." To build that cohesion, Biden's team should be enlisting messengers, including public health experts, "who are trusted by liberals and conservatives," says Ellen Peters, a social scientist at the University of Oregon who has been surveying people's perceptions of the pandemic.

Biden's effort to increase the use of masks could prove an early test of whether his administration can overcome the political and cultural divides in a nation where more than 71 million people voted for Trump. Legal experts have concluded that Biden cannot impose a national mask mandate. He can, however, require the use of masks on federal property and during interstate transportation. But such mandates could flop without a concerted effort to persuade people to embrace masks, says Harvard University health

Published by AAAS

care economist James Stock. If people are ordered to wear masks, they "just won't do it, it's been so politicized."

Biden's history of finding common ground with political opponents could serve him well in this arena, noted Andy Slavitt, head of the Centers for Medicare and Medicaid Services during the Obama administration, during a 5 November presentation at the University of California, San Francisco (UCSF). During the transition, Biden could "lean in hard on this, … doing listening tours" and meeting with governors to firm up support for mitigation measures. "It won't work with all of them, but we don't want to let the perfect be the enemy of the good," adds Eric Toner, an emergency physician and expert on pandemic preparedness at the Johns Hopkins University Center for Health Security.

As president, Biden could use pressure as well as persuasion, by requiring states to follow CDC guidance if they want to qualify for certain kinds of federal funds. "The way the CDC establishes trust and establishes its primacy is to begin to act more aggressively, as long as it has a credible leader," says Robert Wachter, an internist who chairs the UCSF department of medicine. Biden could name his nominees to head key pandemic-fighting agencies, such as CDC and the Department of Health and Human Services, before his inauguration, Ezekiel Emanuel said at the 5 November UCSF event. Emanuel, a University of Pennsylvania oncologist, is now a member of Biden's new task force.

Observers differ on just how much emphasis the Biden administration should place on contact tracing, given that the rapidly rising number of infections could overwhelm any contact tracing system. "Its value is limited unless everything else is going well, too," says Harvard epidemiologist Marc Lipsitch, who compares tracing during a massive outbreak to "trying to clean up a tanker spill with a paper towel."

By the time Biden takes office, many experts fear the United States will be experiencing unprecedented caseloads. Bhadelia says she is particularly frightened by the specter of overwhelmed hospitals and tens of thousands of additional deaths by late January 2021. She predicts "a huge surge in the need for … everything from personal protective equipment to health care workers to medications." Biden has vowed to marshal the needed response, but, Bhadelia says: "The trouble is, we need all those things now."

Even if the Biden administration makes all the right calls once it takes charge of the pandemic fight, many experts expect the battle to be lengthy and grueling. Turning the pandemic around "is gonna be hard," Toner says. "It's gonna be slow." ■

GLOBAL HEALTH

New polio vaccine could boost faltering eradication drive

Sabin's oral vaccine has itself become a source of polio outbreaks. A modern, stabler version may reduce the risk

By **Leslie Roberts**

The Global Polio Eradication Initiative (GPEI) is about to roll out a brand-new vaccine—one that its leaders desperately hope will turn the flagging effort around. If it works as expected, the vaccine just might overcome one of the biggest obstacles to polio eradication: out-of-control outbreaks caused by the polio vaccine itself. If not, GPEI will be back to dousing each outbreak with a vaccine that risks starting another, as eradication slips further from sight.

The new vaccine is so urgently needed that the Bill & Melinda Gates Foundation has paid for nearly 200 million doses to be produced by an Indonesian manufacturer "at risk," even before clinical trials are complete. The World Health Organization (WHO) is ready to grant an emergency use listing, a new approval mechanism never before used for a vaccine, as soon as Indonesia's regulatory authority gives the product the nod. The first drops could be delivered in several countries by the end of the year.

"All the early signs and data look pretty good," says Jay Wenger, who leads polio eradication efforts at the Gates Foundation. But he and others concede that clinical trials to date have involved just 1200 people. The vaccine's full worth won't be known until it is given to millions of children under close scrutiny.

The problem is Albert Sabin's oral polio vaccine (OPV), the workhorse of the eradication program. OPV is cheap, effective, and easy to use, needing just two drops on a child's tongue. It is made from a living virus, attenuated, or weakened, so it can't paralyze but can still replicate in the gut. This means children shed the vaccine virus in their stool for a few weeks after they receive it, spreading it to others who are not vaccinated and protecting them, too. (The killed, injected vaccine adopted by most rich countries protects against paralysis but does not block circulation.) Massive OPV immunization campaigns have driven the wild polio virus almost to extinction; it is now holed up in just Pakistan and Afghanistan.

In places where immunization rates are low, however, OPV can continue to pass among susceptible children for several weeks, mutating as it goes until it has regained its ability to paralyze and spread. This problem is especially acute for the type 2 virus, one of three polio serotypes. Scientists realized long ago that this meant they would eventually have to stop using OPV. But OPV is still the only vaccine that can quash outbreaks in the poor, hot settings where the virus thrives.

> **"Hopefully, we will stop lighting new fires. That is the hope."**
> **Mark Pallansch**, Centers for Disease Control and Prevention

GPEI's answer was to phase OPV out one serotype at a time after each was eradicated in the wild. Type 2, last seen in 1999, would be the first to go. In April 2016, all countries stopped using trivalent OPV and replaced it with bivalent OPV, with the type 2 component removed—a trial run for the eventual cessation of all OPV. The plan was to snuff out the few type 2 outbreaks that would inevitably emerge from the last use of trivalent vaccine with a limited, emergency stockpile of a vaccine effective against just type 2, known as monovalent OPV2 (mOPV2).

The move worked in most places, but mOPV2 spawned new outbreaks in some others. Twenty-three countries, most of them in Africa, are now battling vaccine-derived outbreaks, and the vaccine virus is paralyzing more children than the wild type. The problem is rapidly getting worse, in part because children born after the switch no longer have immunity to type 2. So far this year, more than 600 vaccine-derived cases have been recorded, five times as many as by this point in 2019.

The new vaccine, in the works since 2011 and known as novel OPV2 (nOPV2), is an improved version of Sabin's vaccine, painstakingly engineered to be more genetically stable. The Sabin vaccine was created in the 1940s and '50s by passaging the virus through animal cells until scientists found a suitably weakened form. They now know a few mutations at one end of the viral genome are key to weakening it.

Published by AAAS

To "tighten" or stabilize these Sabin mutations, a global consortium of dozens of researchers introduced 18 nucleotide changes at the so-called 5' end of the viral genome. One change, explains the Gates Foundation's Ananda Bandyopadhyay, who co-chairs the nOPV2 working group at GPEI, was to block a single point mutation that seems to be the "gatekeeper," preceding other events leading to virulence. Other tweaks aim to block another route back to neurovirulence: recombination, in which the Sabin virus swaps genetic material with related viruses and loses the mutations that weaken it.

Data from preclinical research and phase I and II studies in adults, children, and infants in Belgium and Panama are "very, very reassuring," Bandyopadhyay says: The vaccine appears to be as safe and effective as the Sabin vaccine, and so far, extensive sequencing analyses have turned up no evidence of dangerous reversion. Now, the "million-dollar question" is how stable the vaccine will be when used in a much larger population, says WHO's Simona Zipursky, who co-chairs the nOPV2 working group.

A girl receives the oral polio vaccine in Egypt, a candidate country for the introduction of a new vaccine less likely to cause outbreaks.

Wenger hopes it will be hundreds of times more stable. "But [even] a 10-fold increase, which is not much, would be a huge change," he says. "It doesn't have to be absolutely perfect to be useful. It could be enough to let the program catch up."

"I think it is indeed a much, much better vaccine than monovalent OPV2," says WHO's Ondrej Mach, GPEI's team lead for R&D. "But we simply have to be 100% sure." There is a theoretical possibility that the changes aimed at preventing recombination might not work—"a very low probability, high-consequence event" that could undermine faith in the eradication effort, he says. That's why the rollout will be done with "an abundance of caution," Zipursky says, "so if anything unexpected happens, we will be able to identify it and respond to it quickly."

Because the new vaccine will be approved for emergency use only, countries must have an ongoing outbreak. But to meet WHO's 27 strict criteria for the 3-month "initial-use" period, they cannot have used mOPV2 in the same area for the past 12 weeks, because that would make it harder to assess how the vaccine is doing. The problem is that countries don't have the luxury of waiting until the new vaccine is ready; when a new outbreak occurs, they must act immediately with mOPV2. This has resulted in a changing cast of candidate countries as GPEI has waited for WHO's emergency use approval, originally expected this summer.

Countries must also approve emergency use of the vaccine and be able to monitor safety and efficacy, and they have to be able to transport type 2 isolates to a lab for sequencing—fast—to determine whether it arose from the novel vaccine or past use of the monovalent vaccine. Liberia, the Republic of Congo, Sudan, and Egypt are now on the list, says Michel Zaffran, who is stepping down as head of GPEI at the end of the year.

If safety continues to look good and nOPV2 seems to be more stable, "those would be positive signals to roll it out in a bigger way," says John Vertefeuille, who leads polio efforts at the U.S. Centers for Disease Control and Prevention (CDC). GPEI expects to ramp up production to 600 million doses a year.

Even if nOPV2 works as planned, stopping outbreaks will be just as hard as it is with mOPV2, says Mark Pallansch, a polio expert at CDC. It depends on reaching enough children to boost population immunity high enough so the virus can no longer circulate—and countries with ongoing outbreaks have, by definition, been falling short. But even if the new vaccine doesn't quash all vaccine-derived outbreaks, Pallansch says, "Hopefully, we will stop lighting new fires. That is the hope." ■

Leslie Roberts is a journalist in Washington, D.C.

VOICES OF THE PANDEMIC

India's COVID-19 teacher

Kerala health minister K. K. Shailaja has fought the pandemic with preparation, calm, and compassion

By **Vaishnavi Chandrashekhar**

When the World Health Organization (WHO) issued its first statement on the spread of a novel coronavirus in Wuhan, China, on 18 January, few local governments in India paid close attention. But K. K. Shailaja, the diminutive woman running the health ministry in the southern state of Kerala, immediately perked up her ears.

Shailaja knew many students from Kerala were studying at Wuhan University, and she understood the havoc an outbreak could cause. In 2018, during her first stint as a minister, she faced an outbreak of Nipah virus, another deadly pathogen spread from animals to people. "We knew anything could happen at any time," she says.

By 24 January, Shailaja had called a meeting of her rapid response team, set up a control room, and mobilized surveillance teams. On 27 January, the first group of students flew back from Wuhan. Three days later, one of them tested positive for COVID-19, becoming India's first confirmed case.

Kerala was ripe for the spread of the virus, with its large urban population, many residents living abroad (and traveling back and forth), and high influx of migrant laborers. Yet with targeted testing, contact tracing, and isolation measures, the leftist state government brought the number of daily new cases down to almost zero in the first few months, flattening the curve far better than the rest of India. As national lockdown measures eased, infections have risen again, but the state seems prepared to keep things from going out of control. Only 0.36% of confirmed cases have died, a mortality rate that is among the lowest in the world and reflects both Kerala's young population and high-quality health care. "In many ways, [Kerala] got it right," says virologist Shahid Jameel, director of Ashoka University's Trivedi School of Biosciences. "They possi-

PHOTO: GIANLUIGI GUERCIA/AFP VIA GETTY IMAGES

Published by AAAS

bly got it right the most of any Indian state."

Much of the credit goes to Kerala's calm and cheerful health minister, often called "Shailaja Teacher" because of her old job as a high school teacher. Although Kerala benefited from historical advantages including the country's highest literacy rates and arguably its best primary health care system, experts say Shailaja's leadership has been critical. "She listens to people, she visits hospitals privately, she talks to doctors," says K. Srinath Reddy, director of the Public Health Foundation of India. "She comes across as a person who is blessed both in ability and humility."

Shailaja has a passion for science that goes back to her time as a physics and biology teacher in the late 1970s. She and her students would read the science section of local newspapers in class, she recalls. "We would have the most interesting discussions about space, the Moon landing, so many things not on the syllabus," she says. Politics, however, eventually drew her in.

In the 1950s, members of her family joined the growing communist movement. Her grandmother took part in local movements against untouchability (the persecution of caste groups seen as lesser or "impure"), sometimes taking young Shailaja along to tumultuous protests. Shailaja says her grandmother taught her to be brave, and not only in politics. Smallpox was once widespread in Kerala, and the sick were often shunned; many people believed patients were cursed by a goddess. But not her grandmother. She would visit patients in their homes and offer them clean water, good food, and traditional herbal remedies. "She was very bold," Shailaja says. "Everybody should have such a grandmother."

After working her way up to leadership positions in one of the state's communist parties, Shailaja was appointed minister of health and social welfare in 2016. Memories of smallpox may have been in her mind in 2018 as she was grappling with the state's first outbreak of Nipah, a bat-borne virus with a human case fatality rate of 50% to 75%. Ignoring warnings, Shailaja visited the worst hit village to calm residents and explain that, although person-to-person transmission can occur with Nipah, especially in hospitals, the risk is low. Her visit was credited with helping prevent a mass flight from the area.

WHO later concluded that Kerala's early response was improvised and health personnel were inadequately trained. The outbreak was contained, however, by isolating patients and tracing and quarantining more than 2000 contacts. In the end there were 19 confirmed cases and 17 deaths. And Shailaja and her team resolved to be better prepared for the next outbreak. They put in place a raft of measures, including improved surveillance and contact tracing systems, standard operating procedures, and hospital protocols. "It's important to build capacity in peacetime," Jameel says, adding, "They've done well in that."

Shailaja also learned how to deal with the thorny social aspects of an epidemic. In the Nipah outbreak, the government initially cremated the dead—which was unacceptable to Kerala's many Muslims. One relative called Shailaja in tears, and she asked her team to find a solution. Finally, they settled on a burial technique: wrapping the body in airtight plastic and burying it 3 meters deep. "We understood that we have to make our own protocols sometimes," she says.

"Until we get a vaccine, all of us will have to sacrifice some pleasures in our lives."

K. K. Shailaja, Kerala state health minister

Kerala's traditionally strong social services have helped fight the new pandemic. During the national lockdown early this year, Kerala provided migrant laborers with shelter and food to keep them from fleeing back to their home states—and potentially spreading the virus. The state started with some of the best health indicators in the country, including infant mortality rates comparable to those in many wealthier countries. Devolution of power has also strengthened citizens' participation and public communication, Reddy notes. "Despite political polarities, at the [village council] level, there is a great deal of social solidarity," he says.

Shailaja built on those advantages, Reddy says, by engaging scientific advice, generating support across ministries, and communicating with the public. "With a fairly educated and politically agile population, so much depends on gaining citizen trust and cooperation, and she has been able to do that effectively," he says.

Still, she has met with resistance from political rivals seemingly resentful of her success as a woman. One sarcastically called her "Nipah Princess," especially after a film called *Virus* was made about the outbreak with a well-known actor playing Shailaja heroically taking charge. More recently, a politician called her the "COVID Queen." Opposition seems likely to increase with the recent surge in cases.

As India's economy reopened and travel increased, clusters of cases cropped up in parts of Kerala, growing into a surge after the harvest festival in late August. By October, the state was seeing some of the country's largest daily increases in cases. Some observers say the government became lax and did not test enough. Others point to the influx of migrant labor. "A lot of people also returned from the Gulf [countries] after the lockdown lifted," Jameel notes.

Fatigue with the early strict measures may have set in, especially among health workers, and the state's governor, a federal appointee, suggested the good health care and low death rates made people unafraid of the virus. When Shailaja spoke with *Science* in August, she said she was preparing for a possible second wave with an increase in hospital beds and recommendations to renew some lockdown measures. "Until we get a vaccine, all of us will have to sacrifice some pleasures in our lives," she said. More recently she acknowledged that festival gatherings and political protests had contributed to the recent surge, and recommended stricter travel restrictions.

Now, more than ever, Kerala's hard-working health minister will need to draw on all of her abilities—and her grandmother's plucky spirit—to find new strategies against the virus. ■

Science's COVID-19 reporting is supported by the Pulitzer Center and the Heising-Simons Foundation.

Vaishnavi Chandrashekhar is a journalist in Mumbai.

ILLUSTRATION: KATTY HUERTAS

Published by AAAS

COVID-19

Pandemic dooms Danish mink—and mink research

Mass culling ordered after discovery of mutations that may affect vaccine efficacy

By **Christa Lesté-Lasserre**

Jens Malmkvist's life's work came to a dramatic end this week. An ethologist at Aarhus University, Malmkvist studies the behavior and welfare of farmed mink, with the aim of giving them a better life as they are raised for fur. But on 9 and 10 November, all 6350 minks at the Aarhus facility were gassed as part of a nationwide cull ordered by the Danish government.

Denmark is seeking to stop the spread of what it deems a dangerous strain of SARS-CoV-2, the virus that causes COVID-19, that's circulating in mink and infecting humans as well. Scientists say mutations in the virus, described this week in a short report, might reduce the efficacy of COVID-19 vaccines. The cull was a "brave" and "dramatic" step, says Peter Ben Embarek, an expert in food safety and zoonoses at the World Health Organization. Malmkvist says it may be the right decision, but mourns the impact on research and the personal toll on his team. "It's incredibly sad and shocking, just happening overnight like that," he says.

Mink farmer Henrik Nordgaard Hansen killed the entire herd at his farm near Næstved, Denmark, last week.

The move reflects a growing concern about the spread of SARS-CoV-2 in mink, reported in six countries. Four—Spain, Sweden, Italy, and the United States—have responded by culling populations at affected farms. The Netherlands went further by expediting a complete ban on mink farming, originally planned for 2024 and now effective by March 2021. The explosive spread of the virus in mink and several documented cases of mink-to-human transmission made the Dutch government fear farms might become a permanent virus reservoir.

The Danish government has an even more troubling concern. SARS-CoV-2 has remained "quite stable" since its emergence in the human population, Ben Embarek says, but spillovers into animals can trigger mutations as the virus adapts to a new host. That has happened in mink. The report, posted online on 10 November by the Statens Serum Institute, a government lab, reported clusters of genetic mutations in the virus infecting the animals. One strain, now called the ΔFVI-spike mutant, has four mutations in the gene coding for the spike protein, which helps the virus enter host cells.

Experiments with the ΔFVI strain showed plasma from recovered COVID-19 patients or rabbits immunized with the virus did not neutralize it as efficiently as the unmutated virus. Because many COVID-19 vaccines aim to elicit an immune response to the spike protein, the worry is that those vaccines might not work well against the ΔFVI mutant, which so far has been found in 12 people in Denmark as well.

The findings are preliminary, but the government considered them serious enough to order the culling of all 12 million remaining minks in Denmark, the world's largest mink producer. About 80% of the animals were scheduled to meet their end this month anyway—and become fur—but the decision means breeding and laboratory stocks will be destroyed as well. Mink farming will be banned at least until the end of 2021.

Biologist David Kennedy of Pennsylvania State University, University Park, agrees mutations in the spike protein "could completely undermine" vaccine efficacy, especially if they occur in the part of the protein that binds to the human receptor, as one of the four mutations in ΔFVI does. He says more data are needed to document the mutations' impact, but in the meantime it's important to keep the strain from spreading. "If the human-to-human transmission of this variant isn't stopped, stopping further introductions of this variant from the mink won't matter."

Researchers in the Netherlands have also studied SARS-CoV-2 evolution in mink. In a paper published online by *Science* on 10 November, they reported finding the mutation affecting the receptor binding site—but not the other three present in ΔFVI—at four farms. "It did not continue spreading, and we haven't seen it since," says veterinary epidemiologist Arjan Stegeman of Utrecht University. (The team did find that 68% of farm workers and their contacts had antibodies to the virus, underscoring the risk of mink-to-human transmission.)

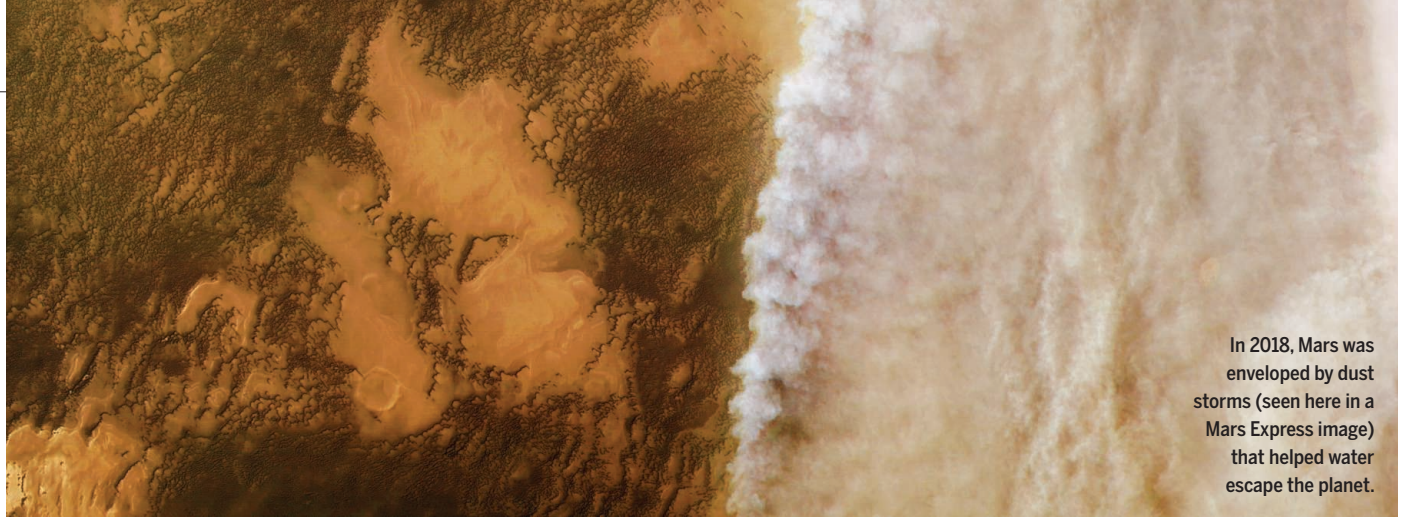
Although mink farming is controversial in Europe, it is likely to survive the pandemic, Malmkvist says, so research into mink welfare remains necessary. His studies have led to improved cage size and design, diets, and bedding materials. Work by his team also influenced Danish legislation that requires breeders to select for less fearful animals and helped establish an animal welfare certification system in the European Union, he says: "I felt part of a team, making a difference."

The fact that his research herd had not seen outbreaks makes the mass slaughter all the more painful. "But we cannot take any risk regarding human health," Malmkvist says. "I must trust that the Danish leadership has made the right decision." ■

Christa Lesté-Lasserre is a freelance journalist in Paris.

PHOTO: MADS CLAUS RASMUSSEN/RITZAU SCANPIX/AFP VIA GETTY IMAGES

Published by AAAS



In 2018, Mars was enveloped by dust storms (seen here in a Mars Express image) that helped water escape the planet.

PHOTO: ESA/DLR/FU BERLIN/CC BY-SA 3.0 IGO

PLANETARY SCIENCE

Dust storms on Mars propel water's escape to space

Orbiters spy potential ongoing planetary "death spiral"

By **Paul Voosen**

Two years ago, Mars went undercover. Martian dust storms are common, but every decade or so, for reasons unknown, a monstrous one goes global, veiling the planet. The storms can be a mortal threat to exploration: The one in 2018 killed off NASA's Opportunity rover by coating its solar panels in dust. But now, researchers say the storms may also be one of the culprits in the ultimate martian cold case: how the once-wet planet lost its water.

Fossilized rivers and deltas etched across Mars suggest water flowed there billions of years ago. Most of it must have somehow escaped to space—yet researchers thought water vapor could not travel high in the frigid, thin atmosphere without condensing into snow and falling back to the surface. New data from NASA's Mars Atmosphere and Volatile Evolution (MAVEN) orbiter, published this week in *Science*, show how churning dust storms may in effect pump water into space. "These escape processes are an effective way to make Mars dry," says Anna Fedorova, a planetary scientist at the Space Research Institute of the Russian Academy of Sciences.

One known escape process comes from the Sun's ultraviolet (UV) light, which can split small amounts of water near the surface of Mars, sending hydrogen and oxygen—both lighter than the planet's mostly carbon dioxide air—percolating to the top of the atmosphere, where they are lost into space. But scientists assume water loss by this mechanism is a trickle.

During the 2018 storm, however, Shane Stone, a graduate student at the University of Arizona, was looking at data from MAVEN, which has studied the planet's upper atmosphere since 2014. One MAVEN instrument directly samples the gossamer atmosphere as the probe dips to its lowest orbital altitude of 150 kilometers, and Stone and his colleagues couldn't believe what it was reporting: While the dust swirled at lower altitudes, a deluge of water was reaching the edge of space. "This was really a smack in the face," Stone says. "The global dust storm stands out in the data like nothing else."

Earlier hints that dust storms might somehow be lofting water came in 2014, when two teams reported on UV observations made in 2007, after the last global dust storm, by the Hubble Space Telescope and the Mars Express orbiter. The teams noticed a fluorescent fog of hydrogen in the upper atmosphere, which faded as the southern hemisphere's summer ended and the storm subsided. The only plausible source for that hydrogen was water. "That was the first hint of something weird going on," says Michael Chaffin, a planetary scientist at the University of Colorado, Boulder, who led the Mars Express work.

Since then researchers using instruments on MAVEN and the European Space Agency's Trace Gas Orbiter (TGO) have found evidence for water high in the atmosphere during the southern hemisphere's summer, when solar heating stirs up dust. That was true even when there wasn't a full-fledged dust storm, says Fedorova, who led the TGO work, published in January in *Science*.

Fedorova and colleagues believe dust drives a positive feedback that pumps water into the upper atmosphere. Sun-warmed dust particles radiate heat into the lower atmosphere, driving winds that kick up more dust, which in turn leads to more heat and more dust. A warmer atmosphere can hold more water vapor, and the dust itself likely drags water along with it as it swirls into the upper atmosphere.

The TGO observations showed water levels gradually increase through the southern summer, but the MAVEN data suggest regional and global dust storms drive the largest spikes. Over the course of 2 days near the start of the 2018 storm, water abundance in the upper atmosphere, normally about 3 parts per million, more than doubled; by summer's peak, the storm and the overall warming ultimately combined to push those levels to 60 parts per million. "It's just a huge influx of water," Stone says.

High in the atmosphere, UV light readily splits the water, allowing the hydrogen and oxygen to be lost to space. But Stone and his colleagues believe another destructive mechanism dominates. Their models suggest carbon dioxide, ionized by particles in the solar wind that bombards the atmosphere, vigorously splits the water molecules apart. "They're making a good case," though the importance of the mechanism needs to be better quantified, says John Clarke, a planetary astronomer at Boston University.

The MAVEN team has calculated that if the observed loss rates persisted throughout martian history, the planet would have lost a global ocean more than 25 meters deep. But how long this dust-driven destruction has operated on Mars is anyone's guess, Stone adds. It could be recent, or it may have been key to drying out the planet billions of years ago. Researchers believe the planet once had a protective magnetic field that failed early in its history, allowing the solar wind to penetrate deeper in the atmosphere, where global dust storms were putting water in harm's way. That could have caused water loss to surge, Chaffin says. "Maybe you fall off a cliff and get this seasonal and dust-driven loss," he says. What MAVEN is witnessing now, he adds, could be the end of a "planetary death spiral." ■

Published by AAAS

RISING FROM THE ASHES

An effort to breed North American ash trees that can resist a deadly beetle could accelerate the return of a forest icon

By **Gabriel Popkin**, *at the U.S. Forestry Sciences Laboratory in Delaware, Ohio; Photography by* **Maddie McGarvey**

On a weekday morning in August, just one pickup truck sat in the sprawling visitors' parking lot here at this U.S. Department of Agriculture (USDA) facility. A decadeslong decline in research funding had been slowly quieting the place—and then came the pandemic.

But in a narrow strip of grass behind a homely, 1960s-era building, forest geneticist Jennifer Koch was overseeing a hive of activity. A team of seven technicians, researchers, and students—each masked and under their own blue pop-up tent—were systematically dissecting 3-meter-tall ash trees in a strange sort of arboreal disassembly line. Over 5 weeks, the researchers would take apart some 400 saplings, peeling wood back layer by layer in search of the maggotlike larvae of the emerald ash borer (*Agrilus planipennis*), the most devastating insect ever to strike a North American tree. Since the Asian beetle was first discovered in Michigan in 2002, it has killed hundreds of millions of ash trees across half the continent and caused tens of billions of dollars of damage.

"We have contests for who can successfully pull out the smallest larvae and the biggest larvae," Koch says. "People get pretty excited and competitive about it. You have to do something, because it is very tedious—and [the larvae] are really gross."

The larvae kill ash trees by burrowing into them to feed on bark and, fatally, the thin, pipelike tissues that transport water and nutrients. They then transform into iridescent green beetles about the size of a grain of rice that fly off to attack other trees. Dead larvae excite Koch and her team the most. Those finds signal trees that, through genetic luck, can kill emerald ash borers, rather than the other way around. Such rare resistant trees could ultimately help Koch achieve her ambitious goal: using time-tested plant-breeding techniques to create ash varieties that can fend off the borer and reclaim their historic place in North American forests.

Forest geneticist Jennifer Koch hopes her ash breeding strategy could save other threatened trees.

Koch focuses mostly on the green ash, one of at least 16 native ash species in North America. "If we don't intervene, there's a good chance that green ash will go extinct," she says. But she is also ramping up work on other ash species; none is safe from the borer. And ultimately, she hopes to expand her approach to other trees brought low by a

Published by AAAS

foreign pest, in a bid to reverse the biological hollowing out of forests set in motion by the transcontinental swapping of species.

Since a devastating fungal blight popped up in the Bronx Zoo in 1904 and went on to kill at least 3 billion chestnut trees, North American forests have been swept by one plague after another, including a fungus that kills elms and an aphidlike insect that kills hemlocks. No tree has come back—but Koch hopes her approach can usher in an unprecedented era of tree revival. "We don't think we have to lose any North American tree species to invasive pests and disease," says plant geneticist Jeanne Romero-Severson of the University of Notre Dame, one of Koch's collaborators.

Not all researchers share such faith in tree breeding. Many researchers believe identifying and releasing natural enemies of the ash borer, an approach called biological control, could help ash sooner. And some wonder whether, even if a better ash emerges, the trees will be able to muscle their way back into profoundly reshaped ecosystems. "It's not like if we have a resistant ash tree, everything is going to be hunky dory," says Deborah McCullough, a forest entomologist at Michigan State University.

Koch, however, feels certain that genes lurking among the billions of ash trees growing across the continent can save the species. "Genetic variation," she says, "is a very powerful survival mechanism."

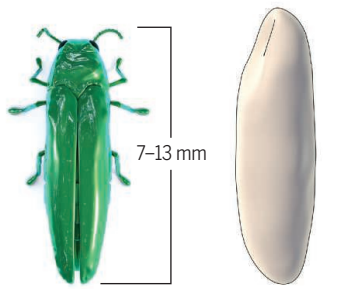
KOCH'S QUEST HAS ITS ORIGINS in a humble place: parking lots around Detroit. That's where, nearly 2 decades ago, experts gathered to examine ash trees that had suddenly died. "It was pretty obvious right away this was not something anybody had ever seen before," recalls McCullough, who cut down some of those trees for study. "We just don't have native insects that do that kind of stuff."

Eventually Eduard Jendek, an entomologist in Bratislava, Slovakia, identified the culprit as the emerald ash borer, which had lived mostly unnoticed in its east Asian homeland. But the insect's arrival in North America set off alarms. Ash is a key genus of temperate hardwood tree. Individuals soar to 35 meters, and species anchor critical ecosystems: black ash in soggy northern wetlands, green ash along Midwestern streams, blue ash in open savannas of Kentucky, white ash in dense mountain forests of Appalachia, and a half-dozen species in the Southwest. Resilient and stress-tolerant, ash was also heavily planted in cities.

Adult borers, scientists soon learned, feed

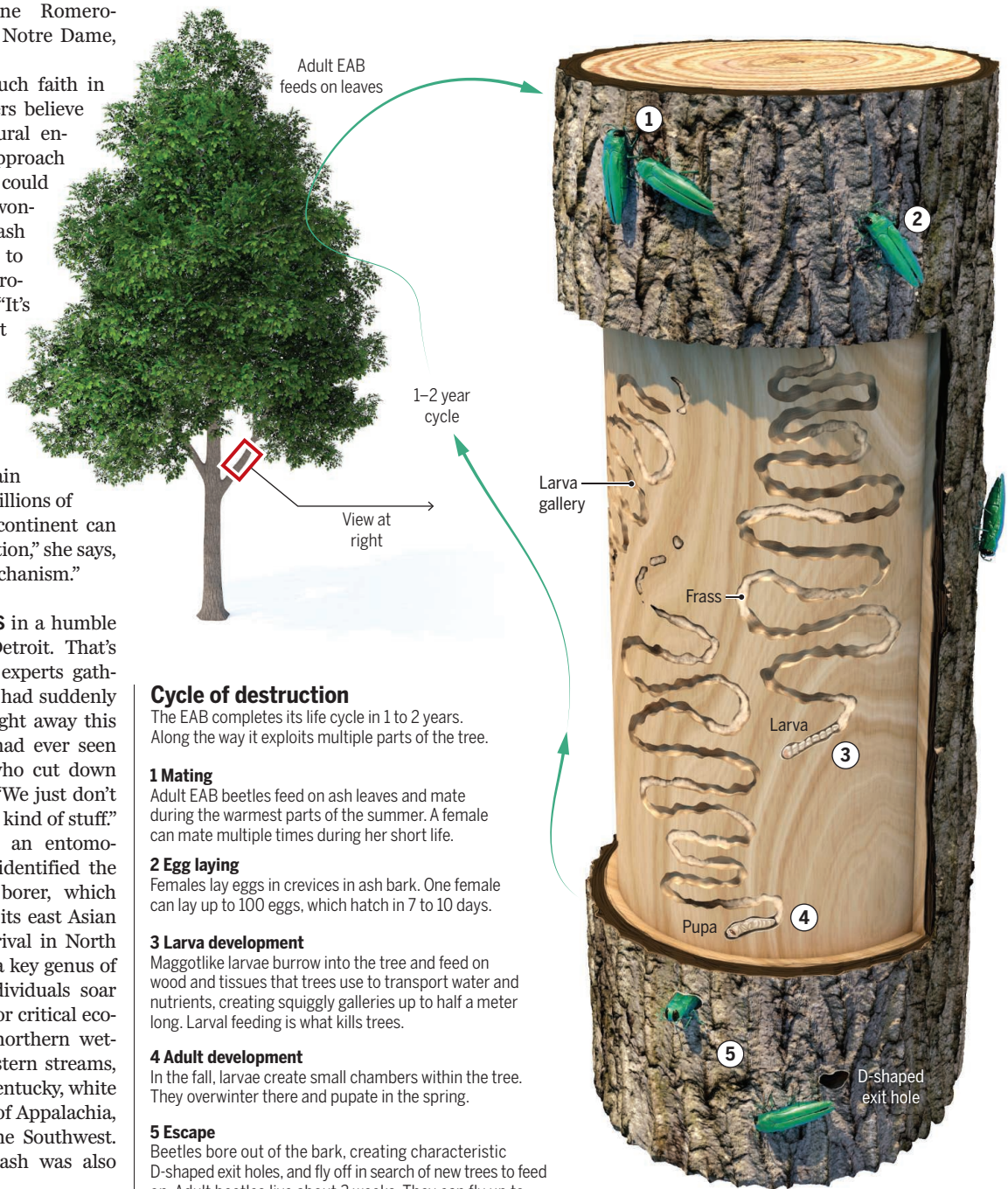
Ash borer invasion

Since slipping into the United States around the turn of the millennium, probably on a container ship, the emerald ash borer (EAB) has spread to 35 states and killed hundreds of millions of trees. Scientists have called it the most destructive forest pest ever introduced to North America.



A cornerstone tree

Comprising at least 16 species native to North America, ash trees are key components of wetland, mountain, and savanna ecosystems. They have also been heavily planted in cities.



Cycle of destruction

The EAB completes its life cycle in 1 to 2 years. Along the way it exploits multiple parts of the tree.

1 Mating
Adult EAB beetles feed on ash leaves and mate during the warmest parts of the summer. A female can mate multiple times during her short life.

2 Egg laying
Females lay eggs in crevices in ash bark. One female can lay up to 100 eggs, which hatch in 7 to 10 days.

3 Larva development
Maggotlike larvae burrow into the tree and feed on wood and tissues that trees use to transport water and nutrients, creating squiggly galleries up to half a meter long. Larval feeding is what kills trees.

4 Adult development
In the fall, larvae create small chambers within the tree. They overwinter there and pupate in the spring.

5 Escape
Beetles bore out of the bark, creating characteristic D-shaped exit holes, and fly off in search of new trees to feed on. Adult beetles live about 3 weeks. They can fly up to 10 kilometers, but usually find a new tree much closer to home.

GRAPHIC: C. BICKEL/*SCIENCE*

Published by AAAS

on ash leaves and lay eggs on ash bark. Burrowing and feeding larvae eventually girdle trees, killing them. North American trees, separated from their Asian cousins by an ocean and millions of years of evolution, had never been exposed to the borer, and lacked chemicals to detect or defeat it. "It's as if they don't even know something is boring into their vascular system and killing them," Romero-Severson says.

Michigan and USDA imposed quarantines on moving ash trees in a bid to contain the pest. But the beetle, an agile flier adept at sniffing out ash trees, slipped through with ease; it has now reached 35 states and Washington, D.C. (see map, right). Many Midwestern city streets became denuded dystopias of dead trees. Some experts predicted the borer could cause the extinction of ash species in North America and began to organize a response.

Koch, who had arrived at the USDA lab in the 1990s as a graduate student and stayed on as a staff scientist, was intrigued. Gregarious and impatient with dogmatic thinking, Koch is more motivated by solving real-world problems than by doing basic research on model organisms. In graduate school, she studied the genetics of air pollution resistance in poplar trees but then moved on to trying to breed resistance to a century-old debilitating bark disease of beech trees caused by a fungus-insect combination.

In 2003, Koch attended an early meeting on the forest pest du jour: the emerald ash borer. During the discussion, she wondered aloud whether some trees might harbor natural resistance to the beetle. "Everyone kind of looked at me like I was nuts," she recalls. "Back then the big headline was: 'Kiss your ash goodbye.' People just didn't believe that there could be resistance within native species."

But Koch persisted. In cities, where the plague first emerged, trees tend to lack diversity, she reasoned; urban foresters tend to plant genetically similar strains widely, making street trees uniquely vulnerable to invasive pests. Trees growing in forests, by contrast, have had millions of years to mutate and mix genes into countless combinations.

The first step, Koch realized, was to survey that genetic diversity before the borer eliminated it. Starting in 2005, ecologist Kathleen Knight of the U.S. Forest Service and Koch set up 150 study plots in infested forests in Ohio, and then repeatedly revisited them, tracing how ash trees in the plots fared. It was "kind of boring" at first, Knight says, because almost all the ash trees simply died. But over time, she found that about one in 1000 green and white ash trees—the most common species in Ohio—produced flushes of new leaves even after borers had killed neighboring trees. Starting in 2008, Knight, Koch, and others cut branches from the biggest and healthiest of those trees, which they dubbed "lingering ash," and grafted them onto root stock of healthy trees. "We try to look for the best of the best," Knight says.

The researchers then developed a way to test for resistance under controlled conditions. Starting in late spring, Koch's team receives coffee filters containing thousands of ash borer eggs reared by Forest Service entomologist Therese Poland in Lansing, Michigan. Researchers place 12 eggs on each stem of hundreds of ash saplings grown from cuttings. Eight weeks later, they dissect the trees and trace the fate of each hatched borer.

David Carey, a technician on Koch's team, was performing that task in August. Sitting under his tent, he snapped green branches off a sapling and saved them for future grafting onto other trees, in case the sapling proved resistant. Then he deftly shaved the thin bark off the trunk, to expose evidence of feeding: "galleries," or winding tunnels bored by the larvae that resemble a kindergartner's marker squiggles.

Soon, Carey found a little yellow head poking out, the larva's body still encased in the tree. He pulled the larva out with tweezers, placed it in a plastic tray with a growing pile of its compatriots, and cataloged its size and location. The tree's failure to kill this larva was a strike against its future in the breeding program. Carey moved down a few centimeters along the tree trunk and started a new dissection.

In 2015, Koch's team published its first major paper on ash, reporting that lingering green ash trees killed significantly more larvae than control trees. That, she says, convinced many who doubted that trees never exposed to an introduced insect could still resist it. But even lingering trees didn't kill enough larvae to save themselves, the researchers found; they still died, just more slowly. So in 2010, Koch started to crossbreed her most resistant trees—the same basic technique used by plant breeders for 10,000 years to produce bigger grains, sweeter fruit, and countless other desired traits. In essence, one of Koch's colleagues joked, they created "TreeHarmony"—a matchmaking service for durable trees. "We're kind of pushing Mother Nature along by putting these trees together," Koch says.

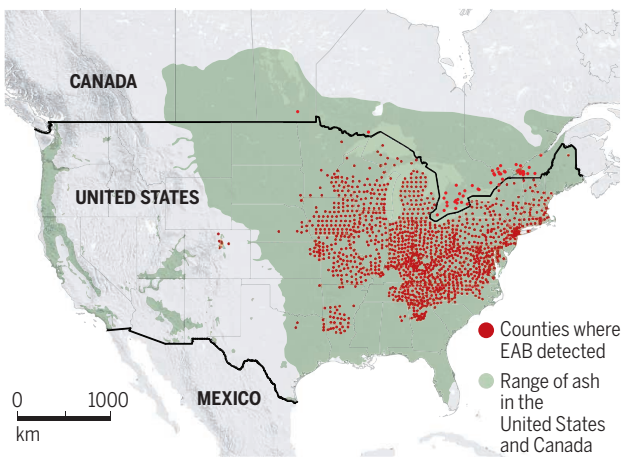
Such traditional breeding is a blunter tool than more fashionable genetic engineering techniques that can precisely transfer or alter single genes. But traits like insect resistance are typically controlled by many genes, not one—and in the case of ash trees, researchers don't know which genes are important. Some may code for chemicals that kill feeding larvae or make wood less digestible; others may make leaves less detectable or palatable to adults. Several generations of breeding and selection will allow these genes to stack up over time, providing ever stronger defense, Koch hopes.

The efforts have borne some fruit. The top-performing offspring of two lingering ash trees kill up to four times as many larvae as their parents, Koch and colleagues have found. "Our best tree so far was 11 out of 11" larvae killed, Koch says; the 12th egg failed to hatch. Now, she's planning to use those top performers to breed an even more resistant generation. If those trees consistently kill 80% to 90% of larvae that attack them in field trials, Koch says they can begin restoration efforts. She hopes restoration plantings could start in about 10 years.

Such a timeline would be fast by tree-breeding standards, but some fear it may be too slow to save the ash. Already, formerly ash-dominated sites now host other trees, grasses, or invasive plants. So, to restore ash, those newcomers would first need to be removed—a potentially expensive, labor-intensive process, McCullough says. "Nature doesn't just sit there and wait for you

A beetle on the march

Since the emerald ash borer (EAB) was discovered in Michigan in 2002, it has spread widely into areas where ash trees grow.



GRAPHIC: A. CUADRA/SCIENCE

Published by AAAS

to come along and plant a tree."

McCullough and others are more bullish on saving existing trees through biological control: identifying, rearing, and releasing insects or other species that kill unwanted pests. The concept has a long history of success in managed landscapes, such as farm fields and orchards. But its record in natural forests is less sterling. For example, biocontrol researchers have spent more than 3 decades seeking and rearing predators to curb the hemlock woolly adelgid, which kills eastern and Carolina hemlock trees; they've made progress but are still far from declaring victory (*Science*, 17 January, p. 238).

Biocontrol researchers say the ash borer might be easier to defeat. In the insect's home range in Asia, researchers have found three species of tiny wasps that lay eggs inside ash borer larvae, and one wasp the size of a grain of sand that parasitizes borer eggs. Parasitoids have often proved to be more successful biocontrol agents than predators, like those that might tame the woolly adelgid, because parasitoids are more likely to target a single species, causing less collateral damage.

In recent years, researchers have released wasps in forests in 30 states where the emerald ash borer is present. Follow-up studies have shown the wasps can find beetles, parasitize them, and reproduce; in some trees, up to 80% of ash borer larvae have wasps living inside them. "There's a lot of data that suggest it's promising," says Poland, who leads the Forest Service's research effort.

Biocontrol is now the "cornerstone" of anti-borer efforts led by the Animal and Plant Health Inspection Service (APHIS), USDA's primary agency for fighting plant pests. Through the end of 2019, APHIS had invested $13.2 million in emerald ash borer biocontrol research, versus $2.4 million in breeding resistant trees. Whereas breeding could pay off eventually, biocontrol may already be protecting regrowing ash trees in some places, says Scott Pfister, director of APHIS's pest management division. "We tend to try to work towards shorter term solutions," he says. "Breeding for resistance is a long-term research effort."

But biological control faces its own limitation: There have to be enough borers in an area to sustain wasp populations. By the time borers have reached such densities, they will have likely already killed almost all large ash trees, experts say.

Larvae of the emerald ash borer chew into trees, where they overwinter beneath the bark before pupating in the spring.

PHOTO: MADDIE MCGARVEY

"Sometimes there's an overexuberance towards the efficacy of biocontrol," says Koch, who feels she has had to meet a higher bar to get her breeding research funded. Still, she hopes the wasps succeed. Effective parasitoids could complement her breeding efforts, she says, by lowering ash borer numbers enough to give partially resistant trees better odds of surviving to reproductive age. That's what happens in Asia, Koch notes, where genes and ash borer enemies conspire to protect trees. The ultimate goal, she says, is "to mimic what happens in nature."

IN THE MEANTIME, Koch worries her funders—primarily USDA—will lose patience with her breeding efforts. So she has embarked on several efforts to speed up the process. In one, she's testing whether ash seedlings grown in containers under artificial light and treated with hormones produce seeds in less than the normal 3 to 4 years. Another seeks to identify the specific compounds that resistant ash trees use to fight the borer. Robert Stanley, a graduate student in Romero-Severson's group, grinds up samples from dissected saplings and analyzes the chemicals in the wood. This is no trivial task; ash, like other trees, is a sophisticated chemical factory, producing hundreds of compounds with structures and functions unknown to science. "Plants are just chemical geniuses," Stanley says.

Most of these compounds probably provide no defense against ash borers. But a handful show up disproportionately in trees that kill the most borer larvae, Stanley has found. And if such compounds reliably predict longer survival in the forest, they could pave the way to rapid testing of wild trees for borer resistance, potentially speeding breeding efforts, Koch says. "We're hoping for some way we can quickly test a tree and say, 'OK, that one's a dog.'"

Despite the progress she's made, Koch says it sometimes "feels like the weight of the world is on us." Hers remains the sole U.S.-based effort to revive ash trees through breeding. (Breeders in the United Kingdom are seeking genetic resistance to a fungus attacking ash trees there.) But she's starting to recruit help by building a network of government and nonprofit organizations to grow and plant resistant ash trees. The collaboration—which includes the conservation organization American Forests; the Holden Arboretum in Kirtland, Ohio; and Fender Musical Instruments, which has long used ash wood for guitars—has plans to establish an ash tree nursery in Detroit. Some of Koch's trees were supposed to be growing there by now, but the shipment was delayed by the pandemic and is now planned for spring.

Ash breeding may be just the beginning. Koch believes her approach could save other North American trees facing introduced threats. Researchers at the University of Rhode Island and North Carolina State University, for example, have reared lingering hemlock trees that may resist the adelgid; Koch hopes to start breeding the best performers. She'd also like to breed beech trees that can beat the bark disease and eventually add butternut, which is threatened by a fungus, as well as elm and chestnut trees. And she knows new diseases and insects will arrive. "My goal," Koch says, "is that when I retire, I leave behind all the tools" that others might need to meet new threats to native trees. ■

Gabriel Popkin is a journalist in Mount Rainier, Maryland.

Published by AAAS

POLICY FORUM

TECHNOLOGY GOVERNANCE: COVID-19

What's next for COVID-19 apps? Governance and oversight

Adaptive governance can help earn social license

By **Alessandro Blasimme** *and* **Effy Vayena**

Many governments have seen digital health technologies as a promising tool to address coronavirus disease 2019 (COVID-19), particularly digital contact tracing (DCT) apps such as Bluetooth-based exposure notification apps that trace proximity to other devices (*1*) and GPS-based apps that collect geolocation data. But deploying these systems is fraught with challenges, and most national DCT apps have not yet had the expected rate of uptake. This can be attributed to a number of uncertainties regarding general awareness of DCT apps, privacy risks, and the actual effectiveness of DCT, as well as public attitudes toward a potentially pervasive form of digital surveillance. DCT thus appears to face a typical social control dilemma. On one hand, pending widespread uptake, assessing DCT effectiveness is extremely difficult; on the other hand, until DCT effectiveness is proven, its widespread use at a population scale is hard to justify. Recognizing that technological uptake is an open-ended process reliant upon social learning and the piecemeal creation of public trust, we suggest that policy-makers set up mechanisms to test effectiveness, oversee the use of DCT apps, monitor public attitudes, and adapt technological design to socially perceived risks and expectations.

To date, both scholarly and policy debates on DCT have largely overlooked the above dilemma, focusing instead on privacy-related issues as the pivotal element of DCT governance (*2*). However, although preserving privacy is of the utmost importance, technical safeguards such as encryption, decentralized data architectures, and temporal limits to data storage have not proved sufficient for DCT apps to quickly diffuse at a population scale.

Social license and trust depend on the capacity of either corporations or governments to meet societal expectations in relation to a specific activity (*3*). Therefore, for DCT to earn social license, such expectations, as well as the factors that cause slow uptake on the part of the public, need to be probed. To increase public trust, the World Health Organization has stressed the importance of

Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, Zurich, Switzerland. E-mail: effy.vayena@hest.ethz.ch; alessandro.blasimme@hest.ethz.ch

ILLUSTRATION: DAVIDE BONAZZI/SALZMANART

Published by AAAS

appropriate oversight for the governance of DCT apps (*4*). Switzerland, for example, has involved the Federal Data Protection and Information Commissioner and the Federal Ethics Committee in the development of the Swiss national DCT app. The French government has sought advice from eight high-profile national expert bodies. Such moves can contribute to the legitimation of a country's approach to DCT. Likewise, oversight mechanisms of a DCT app can play a role in sustaining widespread and continued use by the public.

PUBLIC RELUCTANCE

Studies conducted in April and May 2020 showed that in countries like the United States, Switzerland, and Italy, between 55 and 70% of adults in all age groups were willing to download a contact tracing app (*5*). Yet these figures do not match the current DCT apps uptake. Even in countries with robust privacy safeguards in place, downloads of DCT apps have been below expectations. At the time of writing, the Australian DCT app has been downloaded by 6.5 million (26% of the population), the Italian one by 8 million (13.4%), and the newly released French one by 1.5 million (2.3%). Ireland has about 1.3 million active app users (24%), Switzerland 1.8 million (21.5%), and Germany 16 million (19.3%). As people keep downloading the app, at some point, the desirable number of users may be reached. Decades of research in science and technology studies confirm that such a bell-shaped innovation diffusion pattern is not particularly surprising, as technological uptake does not just rapidly happen by virtue of a technology's presumed usefulness (technological determinism), but owes instead to complex cycles of cultural and political adaptation (social construction of technology) (*6*).

Members of the public cite unauthorized uses of their data beyond COVID-19 containment and access to personal data by IT companies and state authorities as matters of concern (*7*). Moreover, older people and people of lower socioeconomic conditions are considerably less likely to download DCT apps (*8*). Although the public's reservations are understandable, efforts should be made to respond to those concerns and increase the rate of early adoption of DCT systems. Use of opt-out mechanisms rather than opt-in, and large cohort studies in which participants are incentivized to try out the app, could boost initial uptake across demographics. This would help address the dilemma discussed above, leading to a parallel increase in the capacity to assess effectiveness and, at the same time, to exert control over such systems. Failure to do so could lead to premature dismissal of a potentially useful new technology. The Norwegian data protection authority, for instance, stated that the known risks of DCT surveillance outweigh its still unproven public health benefits—a position that caused the Norwegian government to put the system on hold (*9*).

ADAPTIVE GOVERNANCE

When technologies come with known risks but uncertainties about benefits persist, adaptive governance is a valuable policy option. It has a long and respected pedigree—both in academic scholarship and in policy-making—including in areas that resemble severe public health crises, such as natural hazards and disaster risk reduction. In the case of DCT, we know that privacy-related risks are present, alongside risks linked to public surveillance and to technical failure in the presence of a global public health threat. At the same time, DCT effectiveness in containing damage from COVID-19 is still to be assessed.

According to adaptive models, governance should enable social learning and distribute oversight tasks across different actors (*10*). Collaboration between different stakeholders such as developers, health ministries, data protection authorities, experts, and the involvement of lay publics is a key element for an efficient adaptive governance approach (*11*). In the face of the above uncertainties, adaptive governance urges national DCT initiatives to collect and rapidly incorporate new knowledge into their governance.

To effectively implement adaptive governance of DCT, oversight activities should focus on a number of specific adaptive features (*11*, *12*).

Public engagement

Owing to the exceptional circumstances of the COVID-19 crisis, national DCT plans have been rolled out without engaging the public in any phase of the process (*13*). In democratic countries, this is likely to undermine trust in technological solutions, especially if they embody a pervasive surveillance logic that may well appear at odds with democratic ideals. DCT initiatives should thus ensure that they offer regular opportunities of democratic input into the governance of DCT. This can be guaranteed by including lay publics such as civil society representatives, advocacy groups, and nongovernmental organizations in oversight bodies. Moreover, surveys, deliberative forums, and notice-and-comment periods should be regularly offered to increase public input into the governance of national DCT apps. Public engagement should not be seen as a legitimation tool alone, but as a fundamental component of the adaptation process, a precondition for social learning around both anticipated and unanticipated risks. Moreover, public engagement has the potential to mitigate the threat posed by incumbent concentrations of power by state authorities or private companies involved in national DCT strategies.

Technical aspects

The effectiveness of DCT systems in breaking transmission chains should be assessed against previously established public health objectives, such as app penetrance, accuracy, and effectiveness in reducing the health and social burden of the infection. Failure to meet these objectives should lead to reconsidering specific technical aspects of existing DCT strategies.

Regular monitoring of technical parameters about the use and reliability of DCT apps would inform specific strategies to be adopted to increase the rate of downloads and actual use of the apps, and to improve their functioning. Most DCT apps are built with a proactive commitment to privacy-preserving technological features (privacy by design) and only use strictly necessary data (privacy by default). However, no privacy-preserving system is perfect. Oversight bodies should thus regularly test the robustness of adopted privacy-preserving measures and define plans to continuously minimize harms.

Legal aspects

DCT oversight should be able to clarify or, as the case may be, suggest legal definitions for the kind of data collected by DCT apps and the specific roles of all the actors—private or public—involved in development and implementation. In particular, specific types of data like rotating Bluetooth IDs or associated metadata may not have a clear legal status in a given jurisdiction. Ad hoc legislation may also be needed to set specific rules and safeguards around voluntariness and misuses of DCT tools. In Switzerland, for example, such legal provisions were introduced in an amendment to the Epidemics Act before the release of the national DCT app.

Sanctions linked to unlawful handling of personal data are present in most jurisdictions. Increasing public awareness about such legal consequences of data misuse can support trust in DCT systems.

DCT apps operate within national territories. However, cross-border use would facilitate contact tracing while reinstating global mobility. To achieve this objective, technical interoperability and specific legal safeguards about cross-border data

Published by AAAS

exchange must be adopted. The European Commission recently published the "European Interoperability Certificate Governance" specifying technical standards that will enable safe data exchange between national apps (*14*).

Moreover, DCT is not limited to state-sponsored national apps. Private-sector employers and small businesses are already developing their own internal contact tracing systems, and they may make them mandatory for workers and customers. This is happening in the absence of specific legal provisions. DCT oversight bodies should therefore suggest policy guidance to ensure that such private-sector DCT is aligned with constitutional rights and freedoms and will not be used to unduly monitor employees and private citizens. Failure to deploy appropriate regulatory frameworks for private-sector DCT may undermine trust in DCT broadly.

Ethical aspects

If DCT gains traction, ethically complex trade-offs between privacy and effectiveness, or between users' expectations and utility, may need to be addressed. For instance, as new clusters of infection emerge, DCT data may be used to study epidemic dynamics in real time. But this may require lowering privacy safeguards to grant public health authorities access to DCT data. Oversight bodies should thus have monitoring and auditing capacity to ensure that exhaustive information about the scope of data use and data protection safeguards is properly communicated to users through a meaningful electronic informed consent process. Existing guidance on the use of electronic informed consent (*15*) should be adapted to DCT, ensuring that ethical requirements are fulfilled and appropriate ethics review is conducted. Oversight bodies will also have to regularly probe public attitudes and advise policy-makers as to ethically justified, socially accepted, and proportional solutions to such issues.

Notably, DCT runs the risk of exacerbating health inequalities by missing out on people who either do not have a smartphone, have contracts for limited data use, or are not proficient users. Frequently, elders are unfamiliar with advanced smartphone features and may thus be excluded from the potential benefits of DCT—despite representing the most vulnerable social group in terms of COVID-19–related mortality. Furthermore, social groups that are more open to using DCT apps may be taking on a disproportionate burden in making themselves traceable. DCT oversight bodies should be able to monitor these risks and propose, where appropriate, an equitable distribution of the benefits and burdens of DCT. To this aim, it is advisable to include social scientists in oversight bodies, with a mandate to monitor how different social groups respond to and are affected by DCT activities.

REFLEXIVE ADAPTATION

In all of the above domains, oversight bodies should foster reflexive adaptation (*11, 12*) of DCT strategies based on real-world data on actual use of DCT apps. Reflexive adaptation consists in regularly questioning assumptions about design, risks, and users' attitudes to adapt technological features.

One way to proceed is to pay close attention to opportunity costs of new DCT technologies. This implies assessing regularly whether DCT complements or foregoes other containment strategies such as manual tracing methods of established effectiveness—for example, on grounds of representing a cheaper alternative.

Moreover, reflexivity amounts to the capacity to leverage social learning to detect emerging patterns of discrimination and unfair treatment—faced, for instance, by nonusers and people who do not possess the latest smartphone models, or can only afford low–data-use contracts. A further element requiring reflexive capacity is the possible normalization of digital surveillance within and beyond the realm of public health. For instance, DCT apps could be developed to incorporate functions, e.g., QR codes for entry to facilities, that also enable contact tracing—as lately seen in China and the United Kingdom. Unrolling such pervasive forms of control might generate habituation to their use in other domains such as work, schools, public transportation, and so on. Reflexive vigilance of these potential long-term effects is of the utmost importance to prevent the erosion of civil liberties and human rights.

A further hallmark of reflexive adaptation is the capacity to question basic assumptions of DCT models regarding, for instance, users' risk-related behaviors. Although it is generally assumed that DCT alerts are empowering for individuals, different people have different ways of making sense of risk. Absent appropriate user behaviors, the actual effectiveness of DCT will likely be limited. It is therefore important to collect evidence that helps clarify how users act upon being notified by a DCT app. This evidence can be used in efforts aimed at sensitizing users to follow best practices and recommendations about testing and self-isolation.

ROBUST OVERSIGHT

The rapid deployment of DCT apps represents one of the largest experiments in public health surveillance ever attempted—and certainly the first one relying so strongly on digital platforms. We have argued that DCT governance should be focused on evidence collection and planned adaptation to address numerous uncertainties. In the context of a global crisis requiring rapid responses, this approach has two further advantages: It allows governance structures to coevolve with technological solutions while they are already in use, and it can reduce the high cost of intervening in an already widespread technology.

Whatever form, mandate, and composition individual countries will establish, the creation of oversight structures around DCT is of paramount importance and cannot be delayed. Robust oversight will nurture public trust and will contribute to stronger ethical safeguards and to the assessment of DCT's contribution to a safer coexistence with the virus until effective vaccines become available.

COVID-19 found the world unprepared, but now it is time for governments to carefully predispose all the necessary measures to boost resilience and minimize future harms. This model will arguably be useful for other technologies and in case of future large-scale crises—in public health and possibly beyond. ■

REFERENCES AND NOTES

1. I. G. Cohen, L. O. Gostin, D. J. Weitzner, *JAMA* **323**, 2371 (2020).
2. U. Gasser, M. Ienca, J. Scheibner, J. Sleigh, E. Vayena, *Lancet Digital Health* **2**, e425 (2020).
3. J. Morrison, *The Social License* (Palgrave MacMillan, 2014).
4. "Ethical considerations to guide the use of digital proximity tracking technologies for COVID-19 contact tracing" (2020); www.who.int/publications/i/item/WHO-2019-nCoV-Ethics_Contact_tracing_apps-2020.1.
5. E. Hargittai, E. Redmiles, *Sci. Am.* (2020); https://blogs.scientificamerican.com/observations/will-americans-be-willing-to-install-covid-19-tracking-apps/.
6. W. E. Bijker, T. P. Hughes, T. J. Pinch, *The Social Construction of Technological Systems* (MIT Press, 1987).
7. Ipsos MORI, "Survey on COVID-19 track and trace smartphone app for The Health Foundation" (2020); www.ipsos.com/ipsos-mori/en-uk/survey-covid-19-track-and-trace-smartphone-app-health-foundation.
8. Ipsos MORI, "Demographic divide in likelihood to download and report symptoms on the Government's contact tracing app" (2020); www.ipsos.com/ipsos-mori/en-uk/demographic-divide-likelihood-download-and-report-symptoms-governments-coronavirus-contact-tracing-app.
9. G. Ursin, I. Skjesol, J. Tritter, *Health Policy Technol.* (2020). 10.1016/j.hlpt.2020.08.004
10. C. Folke, T. Hahn, P. Olsson, J. Norberg, *Annu. Rev. Environ. Resour.* **30**, 441 (2005).
11. E. Vayena, A. Blasimme, *J. Law Med. Ethics* **46**, 119 (2018).
12. A. Blasimme, E. Vayena, in *Oxford Handbook of Ethics of AI*, F. Pasquale, M. Dubber, S. Das, Eds. (Oxford Univ. Press, 2020), pp. 703–718.
13. M. M. Mello, C. J. Wang, *Science* **368**, 951 (2020).
14. European Commission, European Interoperability Certificate Governance: A security architecture for contact tracing and warning apps; https://ec.europa.eu/health/sites/health/files/ehealth/docs/mobileapps_interop_certificate_governance_en.pdf.
15. *Fed. Regist.* **81**, 90855 (2016).

10.1126/science.abd9006

Published by AAAS

PERSPECTIVES

A 69-year old receives a COVID-19 vaccine in a Phase 3 trial. It is important to understand how different groups respond to COVID-19 vaccines to establish the most effective deployment strategy.

PHOTO: PAUL HENNESSY/NURPHOTO VIA GETTY IMAGES

VIEWPOINT: COVID-19

Understanding COVID-19 vaccine efficacy

Vaccine efficacy in high-risk groups and reduced viral shedding are important for protection

By **Marc Lipsitch**[1] *and* **Natalie E. Dean**[2]

The elderly and people with comorbidities are at greatest risk of severe coronavirus disease 2019 (COVID-19). A safe and effective vaccine could help to protect these groups in two distinct ways: direct protection, where high-risk groups are vaccinated to prevent disease, and indirect protection, where those in contact with high-risk individuals are vaccinated to reduce transmission. Influenza vaccine campaigns initially targeted the elderly, in an effort at direct protection, but more recently have focused on the general population, in part to enhance indirect protection. Because influenza vaccines induce weaker, shorter-lived immune responses in the elderly than in young adults, increasing indirect protection may be a more effective strategy. It is unknown whether the same is true for COVID-19 vaccines.

For COVID-19, age-structured mathematical models with realistic contact patterns are being used to explore different vaccination plans (*1*, *2*), with the recognition that vaccine doses may be limited at first and so should be deployed strategically. But as supplies grow large enough to contemplate an indirect protection strategy, the recommendations of these models depend on the details of how, and how well, these vaccines work and in which groups of people. How can the evidence needed to inform strategic decisions be generated for COVID-19 vaccines?

Phase 3 vaccine trials are designed to assess individual-level efficacy and safety. These trials typically focus on a primary endpoint of virologically confirmed, symptomatic disease to capture the direct benefit of the vaccine that forms the basis for regulatory decisions. Secondary endpoints, such as infection or viral shedding, provide supporting data, along with analyses of vaccine efficacy in subgroups. Nonetheless, unanswered questions about COVID-19 vaccine characteristics are likely to remain even after trials are completed. First, trials are typically not powered to establish subgroup-specific efficacy, yet the performance of the vaccine in high-risk groups affects the success of a direct-protection strategy. Second, can vaccines prevent infection or reduce contagiousness? This matters for achieving indirect protection. Expanding ongoing efforts or planning new studies may generate the data needed to address these questions.

For estimating subgroup-specific efficacy, randomized controlled trials can provide early estimates, yet these will have wide confidence intervals, leaving substantial uncertainty about true effects in high-risk subgroups. This uncertainty would be greater in interim analyses that are based on the number of events across the whole trial population and may be exacerbated if high-risk participants are more cautious and have lower exposure to infection, reducing their contribution to the efficacy estimates.

There are several strategies to address subgroup-specific efficacy, some of them already in place. Ensuring that high-risk adults are well represented in the trial population can be achieved by setting minimum enrollment targets for older adults and/or adults with comorbidities. Another consideration relates to the stopping rules for interim analyses in trials. Vaccine trials with early interim analyses that are planning to discontinue randomization and vaccinate placebo participants after declaring efficacy are most prone to subgroup uncertainty. To improve the precision of efficacy estimates in high-risk subgroups, regulators could insist that interim analyses be performed only after a certain number of confirmed disease cases occur in these subgroups, in addition to existing monitoring of the overall number of events in the study.

Trials that maintain blinded follow-up to assess long-term efficacy and safety may also generate more-reliable evidence on age-specific effects. For example, the World Health Organization's Solidarity Vaccines Trial will preserve placebo-controlled follow-up through month 12 or when an effective vaccine is deployed locally (*3*). However, depending on where the trials are being done and whether the vaccine becomes rapidly available in sufficient quantities after emergency-use authorization in the population undergoing the trial, it may become unethical and/or impractical to ask participants in some subgroups to forego access to an available vaccine. For vaccine candidates evaluated in multiple trials, such as the Oxford-AstraZeneca vaccine being studied in the United Kingdom, South Africa, Brazil, and the United States, meta-analyses can synthesize results across locations to improve precision of subgroup-specific effect estimates.

Ideally, the phase 3 trials in progress will identify more than one safe, effective vac-

[1]Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. [2]Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA. Email: mlipsitc@hsph.harvard.edu; nataliedean@ufl.edu

Published by AAAS

cine for regulatory approval and deployment. Postapproval studies will then take on an important role for continued assessment of vaccine effectiveness. These may include individual- or community-level randomized trials to compare different active vaccines without a control arm, as in the U.S. Department of Defense's individually randomized Pragmatic Assessment of Influenza Vaccine Effectiveness in the DoD (PAIVED) trial, which assesses the relative merits of three licensed influenza vaccines (NCT03734237).

Another approach to amass evidence on subgroup-specific efficacy is postapproval observational studies. This includes active surveillance of high-priority cohorts from, for example, nursing homes or assisted living facilities, as has been done for influenza. This also includes test-negative designs, which are routinely used to assess vaccine effectiveness (*4*). Symptomatic individuals that test negative for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) function as controls for test-positive cases, and their vaccination status is compared, adjusting for selected confounders. Test-negative designs can be integrated into outpatient testing in the community (*5*) or use emergency department visits to estimate vaccine effectiveness against severe disease (*6*). To rapidly establish these systems, researchers can leverage ongoing influenza surveillance. Conveniently, these programs can simultaneously monitor more than one vaccine, enabling assessment of their relative merits.

A key limitation of observational studies is confounding. There may be many differences between individuals who do and do not get vaccinated, which may create noncausal correlations between vaccine status and outcomes. Although such biases can threaten any observational study of vaccine effectiveness, there are some approaches to detect such biases and reduce their magnitude (*7*, *8*).

The clearest evidence of indirect protection is from a vaccine that prevents infection entirely, thereby reducing transmission. These data will be generated in efficacy trials that include infection as a secondary endpoint. This endpoint is measured by a specialized assay to distinguish an infection-induced response from a vaccine-induced antibody response. A vaccine can provide indirect protection even if it does not fully prevent infection (see the figure). Vaccines that reduce disease severity can also reduce infectiousness by reducing viral shedding and/or symptoms that increase viral spread (e.g., coughing and sneezing). A worst-case scenario is a vaccine that reduces disease while permitting viral shedding; this could fail to reduce transmission or conceivably even increase transmission if it suppressed symptoms.

To assess a vaccine's impact on infectiousness, some phase 3 trials examine the amount or duration of viral shedding in laboratory-confirmed, symptomatic participants by home collection of saliva samples and frequent polymerase chain reaction (PCR) testing. However, this would not capture any change in viral shedding for asymptomatic participants. Moreover, serology tests detect previous infection and cannot reconstruct shedding during active infection. To measure viral load in both symptomatic and asymptomatic participants, it is necessary to conduct frequent (e.g., weekly) viral testing, irrespective of symptoms, to capture participants during their period of acute infectiousness. The Oxford-AstraZeneca vaccine trial is testing participants in the United Kingdom for the virus weekly regardless of symptoms, but not in other trials for which protocols have been released. Even weekly testing will not give detailed information about the effect of the vaccine on viral shedding, and the relationship between viral loads and infectiousness is unknown; nonetheless, this approach is likely to provide some evidence if viral loads are on average lower among vaccinated people. Human challenge vaccine studies, in which individuals in a randomized controlled trial are deliberately exposed to the virus, could generate high-quality data on the effect of vaccines on viral shedding (*9*).

Other approaches exist to directly estimate infectiousness without the need to extrapolate from viral load. Add-on household studies can supplement efficacy trials. Investigators can follow household members or other close contacts of infected participants to assess the vaccine's effect on infectiousness, as has been implemented for the respiratory disease pertussis (also called whooping cough) (*10*). Viral sequencing could be used within the trial to link infector-infectee pairs and better estimate indirect effects (*11*). Another strategy is to design cluster-randomized trials in which indirect effectiveness is a primary outcome. In influenza vaccine trials, health care workers at nursing homes were cluster-randomized to be offered vaccine or not, and the endpoints were mortality, influenza-like illness, or influenza infection in the patients they cared for (*12*).

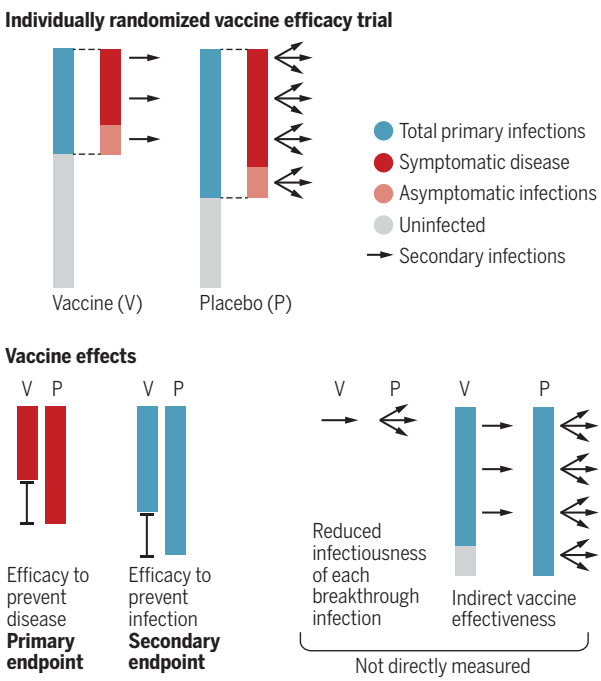
Observational studies may also be helpful, but, in general, measuring indirect effects of vaccines is even harder than detecting direct effects. It is urgent, therefore, to obtain evidence on how each candidate vaccine affects infectiousness either before approval or soon after, when scarcity may justify randomized distribution of a vaccine.

Other open questions about the rapidly developed COVID-19 vaccines include long-term safety (indicating the critical need for pharmacovigilance activities), the duration of vaccine protection, the efficacy of a partial vaccination series or of lower doses (*13*), the vaccine's level of protection against severe infection and death, efficacy by baseline serostatus, and the potential for the virus to evolve to escape vaccine-induced immunity. The answers to such questions inform the optimal use of any vaccine.

Availability of a COVID-19 vaccine will initially be limited, and so several expert

Vaccine effects

Vaccines provide direct protection by reducing susceptibility to disease or infection. Vaccines provide indirect protection by reducing the number of people infected in a population or their infectiousness. These vaccine effects can be assessed in clinical trials by measuring the efficacy to prevent disease, to prevent infection, and to reduce infectiousness, as well as in studies to assess indirect effects of the vaccine (*15*).



GRAPHIC: KELLIE HOLOSKI/*SCIENCE*

Published by AAAS

committees are exploring strategic prioritization plans. Health care workers are a common first-tier group (*14*), which in turn preserves health care systems by protecting those who run them. A next priority is to directly protect those who are at highest risk of death or hospitalization when infected: specifically, those over 65 and people with certain comorbid conditions. This strategy may be optimal for reducing mortality even if the vaccine is somewhat less effective in these groups (*2*). But if a vaccine offers little to no protection in high-risk groups yet is able to reduce infection or infectiousness in younger adults, an indirect strategy could be preferred as vaccine supplies become large enough (*1*, *2*). A worst-case scenario for an effective vaccine is one that reduces disease in younger adults but provides neither direct nor indirect protection to high-risk groups, leaving the most vulnerable at risk. Knowing these vaccine characteristics is important when evaluating the relative merits of other products. Fortunately, there are many vaccine candidates in development that use a mixture of innovative and existing technologies. Although vaccines may vary in their characteristics, having reliable evidence on direct and indirect protection can help plan how to use these vaccines in a coordinated way. ■

REFERENCES AND NOTES

1. L. Matrajt, J. Eaton, T. Leung, E. R. Brown, *medRxiv* 10.1101/2020.08.14.20175257 (2020).
2. K. M. Bubar *et al.*, *medRxiv* 10.1101/2020.09.08.20190629 (2020).
3. P. Krause *et al.*, *Lancet* 10.1016/s0140-6736(20)31821-3 (2020).
4. H. Chua *et al.*, *Epidemiology* **31**, 43 (2020).
5. Centers for Disease Control and Prevention (CDC), US Flu VE Network (2020); www.cdc.gov/flu/vaccines-work/us-flu-ve-network.htm.
6. Q. Chen *et al.*, *J. Infect. Dis.* **211**, 1045 (2015).
7. M. Lipsitch, A. Jha, L. Simonsen, *Int. J. Epidemiol.* **45**, 2060 (2016).
8. N. E. Dean, M. E. Halloran, I. M. Longini Jr., *Am. J. Epidemiol.* 10.1093/aje/kwaa084 (2020).
9. S. K. Shah *et al.*, *Science* **368**, 832 (2020).
10. M. E. Halloran *et al.*, *J. Am. Stat. Assoc.* **98**, 38 (2003).
11. R. Kahn *et al.*, *medRxiv* 10.1101/2020.09.14.20193789 (2020).
12. W. F. Carman *et al.*, *Lancet* **355**, 93 (2000).
13. S. Riley, J. T. Wu, G. M. Leung, *PLOS Med.* **4**, e218 (2007).
14. National Academies of Sciences, Engineering, Medicine, "Framework for equitable allocation of COVID-19 vaccine" (Consensus Study Report, The National Academies Press, Washington, DC, 2020).
15. M. E. Halloran, I. M. Longini Jr., C. J. Struchiner, *Design and Analysis of Vaccine Studies* (Springer, 2010).

ACKNOWLEDGMENTS

M.L. receives funding from cooperative agreement 1U01CA261277 from the U.S. National Institutes of Health. N.E.D. receives funding from NIH/NIAID R01-AI139761. We thank R. Venkayya and N. Grassly for helpful comments. M.L. receives honoraria and consulting fees from Merck, Affinivax, Sanofi Pasteur, and Antigen Discovery; receives research funding (institutional) from Pfizer; and provides unpaid scientific advice to Janssen, Astra-Zeneca, and Covaxx (United Biomedical).

Published online 21 October 2020
10.1126/science.abe5938

CORONAVIRUS

Enhancing host cell infection by SARS-CoV-2

Neuropilin-1 binds the furin-processed spike protein of SARS-CoV-2 to promote virus entry

By **Margaret Kielian**

The current global pandemic of coronavirus disease 2019 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A critical initial step of infection is the interaction of the virus with receptors on host cells. In the case of SARS-CoV-2 and other coronaviruses, this receptor binding occurs through the spike (S) protein on the virus surface. Both SARS-CoV-2 and the related SARS-CoV, which caused an outbreak in 2003, bind to angiotensin-converting enzyme 2 (ACE2) on human cells. However, the observed differences in tissues that are infected by these two viruses (tropism) suggests that additional host factors may be involved. On pages 861 and 856 of this issue, Daly *et al.* (*1*) and Cantuti-Castelvetri *et al.* (*2*), respectively, show that the membrane protein neuropilin-1 (NRP1) promotes SARS-CoV-2 entry and explain how NRP1 interacts with the SARS-CoV-2 S protein. The results suggest the S protein–NRP1 interaction as a potential antiviral target.

Coronaviruses are enveloped RNA viruses that can cause human diseases that range from the common cold to severe and fatal respiratory infections. It is thought that both SARS-CoV-2 and SARS-CoV bind to ACE2 on the host cell surface, are internalized by endocytosis, and fuse with the endolysosome membrane to deliver the viral genome for replication in the host cell (*3*, *4*) (see the figure). The viral S protein mediates this key membrane fusion reaction, but its activity requires several processing steps. S is synthesized as a large membrane protein that is cleaved into two components, S1 and S2, which remain noncovalently associated (*5*, *6*). Cleavage is required for infection and can occur during virus particle production or virus entry into the target cell. The S1 protein forms the "head" of the molecule and mediates binding to ACE2. The S2 protein is anchored in the virus membrane and mediates membrane fusion. S2, like the fusion proteins of influenza virus and HIV-1, inserts a hydrophobic fusion peptide at its amino terminus into the cell membrane and then folds back to merge the host and virus membranes (*7*). The S2 protein needs a further proteolytic step to "liberate" its fusion peptide, and this is carried out by transmembrane protease serine 2 (TMPRSS2) or other proteases (*3*).

A potentially important difference between SARS-CoV-2 and SARS-CoV is the mechanism of S protein cleavage into S1 and S2. In SARS-CoV, this is caused by host cell proteases called cathepsins, which are located within endocytic compartments. However, the sequence of the SARS-CoV-2 S protein contains a series of basic amino acids at the S1-S2 junction. Such polybasic sites can be substrates for furin, a protease that is present in the secretory pathway and endocytic compartments (*8*). Studies with SARS-CoV-2 show that its S protein is cleaved by furin during virus production and that this cleavage promotes subsequent virus infection (*3*, *6*). Thus, a notable difference between the S proteins of these two coronaviruses is the protease that carries out the S1-S2 cleavage reaction. Furin cleavage also produces a potentially important remnant: the polybasic site that remains at the carboxyl terminus of SARS-CoV-2 S1 after cleavage.

Neuropilins are a family of membrane proteins that were originally identified because of their involvement in angiogenesis (blood vessel formation) and axon guidance (*9*). The neuropilins are co-receptors for molecules such as vascular endothelial growth factors (VEGFs) and semaphorins, and recent studies have demonstrated their upregulation during tumor angiogenesis and their potential as anticancer targets. Both NRP1 and NRP2 can bind the carboxyl-terminal sequences generated by furin processing of molecules such as VEGFs, with the sequences fitting into a pocket on the b1 domain of the NRP (*9*). Detailed studies of such NRP1-peptide interactions show that binding to the b1 pocket requires the sequence Arg/Lys-X-X-Arg/Lys (R/K-XX-R/K, where X can be any amino acid) at the carboxyl terminus of the protein or peptide (*10*). This "C-end rule," or CendR, can thus predict whether a protein is a candidate for binding to NRPs.

Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
Email: margaret.kielian@einsteinmed.org

Published by AAAS

Daly *et al.* and Cantuti-Castelvetri *et al.* found that the sequence of the S1-S2 junction of virus isolates from human patients suggested that they fit the C-end rule, with Arg-Arg-Ala-Arg (RRAR) predicted to form the carboxyl-terminal sequence of the furin-cleaved S1. They showed that NRP1 promoted infection of human cell lines by SARS-CoV-2 and by lentivirus pseudotypes that contained the SARS-CoV-2 S protein on their surface. NRP1 was not the only host factor that promoted SARS-CoV-2 infection, but even when both ACE2 and TMPRSS2 were present, NRP1 gave an additional increase. This was due to an increase in virus uptake into the cell rather than an increase in virus binding to the cell surface. The promotion of virus infection by NRP1 was inhibited by the addition of a soluble NRP1 or by an antibody that mapped to the binding pocket on NRP1. Further analyses revealed that S1 or its carboxyl-terminal region interact with NRP1, and this was inhibited by mutations in the NRP b1 pocket. SARS-CoV-2 mutants in which the polybasic cleavage site was deleted or S was made resistant to furin cleavage were insensitive to NRP1 expression.

Daly *et al.* showed that a peptide derived from the S1 carboxyl terminus binds to the NRP1 b1 domain with micromolar affinity. They solved the crystal structure of the complex, which revealed that the peptide is positioned in the b1 pocket, similar to a VEGF peptide that was crystallized with b1 (*9*). A small-molecule antagonist of NRP1 that inhibits VEGF binding (*11*) also inhibited the b1-S1 peptide interaction and virus infection.

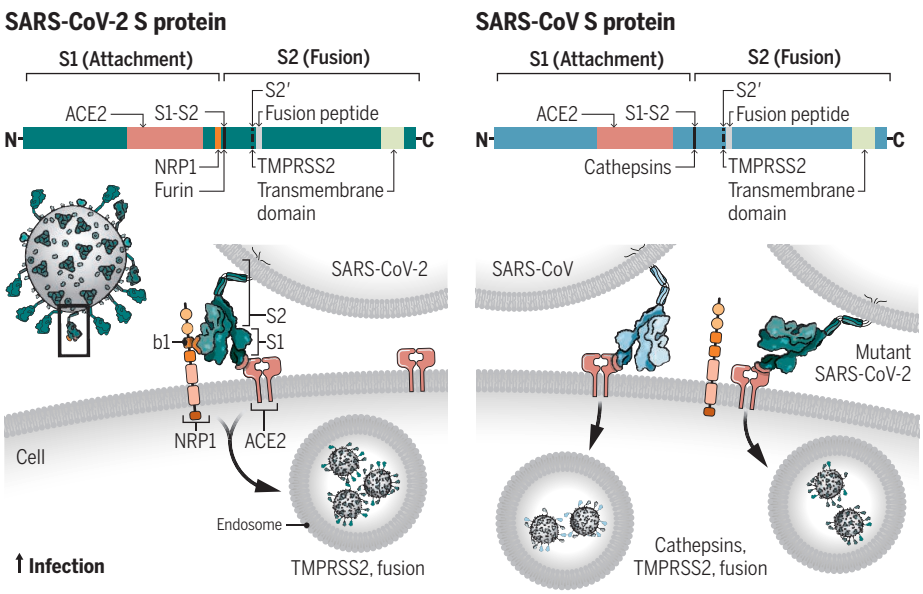
C-end rule peptides had previously been shown to mediate uptake of particles or bacteriophages by cells and tissues (*10*). Cantuti-Castelvetri *et al.* conjugated the S1 peptide onto nanoparticles and administered them intranasally to mice. The mouse olfactory epithelium expresses NRP1, and the authors observed significant uptake of the peptide-conjugated particles in this site. Similar to previous results (*10*), the S1 peptide–coated particles were observed to also travel into the neurons and blood vessels of the cortex. To address the possible role of NRP1 in human SARS-CoV-2 infection, the authors used available data to confirm expression of both *NRP1* and *NRP2* RNA in human lung and olfactory epithelial tissue. They also showed that five out of six autopsy samples of olfactory epithelia from human COVID-19 patients were positive for both S protein and NRP1. These results from mice and humans are intriguing given that many COVID-19 patients lose their sense of smell.

Together, these papers establish NRP1 as a host factor for SARS-CoV-2 and suggest interesting parallels and differences with the roles of NRPs in infection by other viruses. For example, both the human T cell lymphotropic virus SU protein and the Epstein Barr virus (EBV) gB protein are processed by furin, and infection by these viruses is promoted by NRP1 (*12*, *13*). EBV, which infects nasopharyngeal epithelial cells, shows an apparent reciprocal effect in which NRP1 promotes infection, whereas NRP2 inhibits it (*13*). Although Daly *et al.* showed that S1 protein can also bind NRP2, its role in SARS-CoV-2 infection is unknown. NRP2 is a receptor for Lujo virus, but this does not involve binding to the b1 pocket (*14*). These and other examples indicate that a number of viruses have evolved to use NRPs during infection, but much remains unclear.

There is more to learn about the promotion of SARS-CoV-2 infection by NRP1. The virus can propagate in the absence of furin cleavage in some cultured cells, where cathepsin cleavage may be sufficient, but in vivo, the virus relies on furin processing, which enhances viral pathogenesis in a hamster model (*15*). How does NRP1 binding affect the virus internalization pathway, and does it act as a co-receptor with ACE2? C-end rule peptide binding to NRP1 can promote vascular leakage and tissue penetration, especially when the peptides are presented on a multivalent particle (*10*). Does NRP1 similarly help to promote SARS-CoV-2 dissemination and spread? Together, Daly *et al.* and Cantuti-Castelvetri *et al.* show that furin processing, which has an important role in the maturation and fusion activity of the SARS-CoV-2 S protein, also acts to generate ligands on the virus particle that bind NRP1. The availability of small molecule inhibitors of the NRP1-C-end rule peptide interaction suggests a potential antiviral strategy. Defining the importance of this interaction in vivo will be a vital next step. ■

Model for SARS-CoV-2 processing and entry

Proteolytic processing of SARS-CoV and SARS-CoV-2 S proteins facilitates virus entry. SARS-CoV and SARS-CoV-2 bind to ACE2 at a region on S1. Furin cleavage at the S1-S2 junction exposes the C-end rule peptide on SARS-CoV-2 S1 and allows binding to NRP1. Subsequent processing by cathepsins and TMPRSS2 allows S2 fusion peptide–mediated membrane insertion and merging of membranes. The absence of a furin cleavage site in SARS-CoV S1 and a SARS-CoV-2 S1 mutant prevents binding to NRP1 and limits virus entry and infection.



ACE2, angiotensin-converting enzyme 2; NRP1, neuropilin 1; S, spike; SARS-CoV, severe acute respiratory syndrome coronavirus; TMPRSS2, transmembrane protease serine 2.

REFERENCES AND NOTES

1. J. L. Daly *et al.*, *Science* **370**, 861 (2020).
2. L. Cantuti-Castelvetri *et al.*, *Science* **370**, 856 (2020).
3. M. Hoffmann *et al.*, *Cell* **181**, 271 (2020).
4. X. Ou *et al.*, *Nat. Commun.* **11**, 1620 (2020).
5. D. Wrapp *et al.*, *Science* **367**, 1260 (2020).
6. A. C. Walls *et al.*, *Cell* **181**, 281 (2020).
7. S. C. Harrison, *Virology* **479-480C**, 498 (2015).
8. S. S. Molloy *et al.*, *Trends Cell Biol.* **9**, 28 (1999).
9. H. F. Guo, C. W. Vander Kooi, *J. Biol. Chem.* **290**, 29120 (2015).
10. T. Teesalu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16157 (2009).
11. A. Jarvis *et al.*, *J. Med. Chem.* **53**, 2215 (2010).
12. S. Lambert *et al.*, *Blood* **113**, 5176 (2009).
13. H. B. Wang *et al.*, *Nat. Commun.* **6**, 6240 (2015).
14. M. Raaben *et al.*, *Cell Host Microbe* **22**, 688 (2017).
15. S. Y. Lau *et al.*, *Emerg. Microbes Infect.* **9**, 837 (2020).

ACKNOWLEDGMENTS

The author is supported by National Institutes of Health (R01-GM057454 and R01-AI075647).

10.1126/science.abf0732

GRAPHIC: V. ALTOUNIAN/*SCIENCE*

Published by AAAS

CHEMISTRY

Peptide synthesis at the origin of life

Small-molecule organocatalysis might have driven the emergence of peptide biochemistry

By **Kamila B. Muchowska** *and* **Joseph Moran**

The chemical origin of life is full of chicken-and-egg conundrums. Among these is the origin of protein synthesis. Nature's protein-based enzyme catalysts are built from the polymerization of amino acids, yet this process itself requires enzymes, adenosine triphosphate (ATP), and, most often, a ribosome. How were the first proteins formed on the path from chemistry to life? Despite more than 65 years of research into the chemical origins of life, there is still no clear answer (*1*). On page 865 of this issue, Foden *et al.* (*2*) tackle the protein-synthesis conundrum from a new angle.

Researchers who study prebiotic chemistry have largely pursued two general approaches to make amino acids polymerize in the absence of biological polymers and ATP. Because the direct polymerization of amino acids requires the loss of water, the simplest strategy is to drive the process with wet-dry cycles (*3*). Although drying processes are plausible on a lifeless Earth, they are incompatible with other chemistries in a system containing more than just amino acids. To make matters worse, polymerization under wet-dry cycles might be partially reversible, thus scrambling information contained in the specific peptide sequences (*4*). Retaining intact information would have been essential to life's further complexification. A second major approach uses chemical condensing agents to drive amino acid polymerization. Some of these condensing agents are known from synthetic chemistry but have no clear prebiotic source. Others are plausible in the early Earth environment, but it is difficult to envision them being produced continuously in the needed quantities (*1*).

It is unsurprising, however, that nature provides a guide to its own beginnings (*5*). In biochemistry, most amino acid polymerization occurs by translation, which requires ribosomes—a massive assembly of proteins and RNA. Building a primordial ribosome under the conditions of a lifeless planet is a hefty and, as of now, unmet challenge. However, microorganisms use biological nonribosomal peptide synthesis for the production of peptide-based natural products. The nonribosomal peptide synthetases that perform this reaction are chemically much simpler than ribosomes.

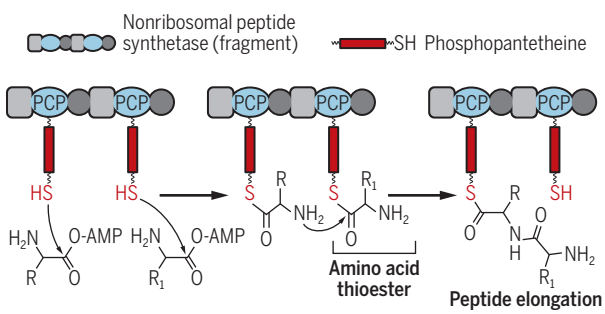
Nonribosomal peptide synthesis begins with the activation of an amino acid's carboxy terminus by ATP-driven adenylation (see the figure). This is followed by nucleophilic attack on the adenylated amino acid by the thiol moiety of the phosphopantetheine prosthetic group of a peptidyl carrier protein (PCP). The resulting PCP-bound amino acid thioesters undergo peptide bond–forming reactions with one another, leading to amino-to-carboxyl-terminal peptide chain growth. The thiol moiety of phosphopantetheine is supplied by cysteine, which is also the prevalent organic source of sulfide in biochemistry (*6*). However, most scientists do not believe that cysteine commonly participated in prebiotic chemistry at its earliest stages. Cysteine is often regarded as a late evolutionary addition to the proteinogenic amino acid pool (*7*, *8*).

Devising a plausible prebiotic route to cysteine has proven to be challenging (*9*, *10*). In biology, cysteine synthesis starts from serine and proceeds through a pyridoxal phosphate cofactor-bound dehydroalanine, an intermediate that has a markedly short lifetime in the absence of the cofactor (*6*). Foden *et al.* proposed how this inherent instability might be overcome by beginning the prebiotic synthesis of cysteine from a nonbiological nitrile analog of the amino acid serine. Proceeding through a doubly acylated dehydroalanine, the acetylated cysteinyl nitrile was obtained in near-neutral water. Given the centrality of cysteine in sulfur biochemistry, the authors wondered if a simpler prebiotic version of nonribosomal peptide synthesis could have relied on this amino acid instead of its more complex biosynthetic successor, phosphopantetheine. Methods that use cysteine moieties in peptide synthesis include cysteinyl thioester–based native chemical ligation (*11*). Although popular, these methods start with a carboxy-terminal amino acid thioester. To make use of this strategy in a prebiotic context, a robust route to amino acid thioesters would be needed—one that could operate without the full toolkit of modern synthetic chemistry (*12*).

A previously reported alternative approach demonstrated sulfide-mediated oligomerization of amino nitriles to yield peptides in water. The process began with amino nitriles, but under conditions of alternating reducing and oxidizing environments and ex-

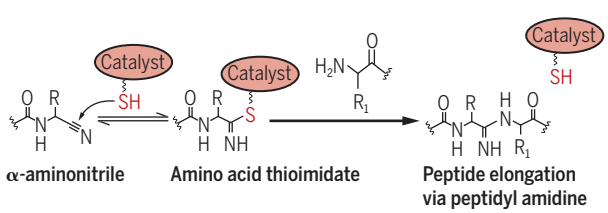
Building proteins in a prebiotic world

Two nonribosomal peptide synthesis reactions highlight the pivotal roles that small-molecule organocatalysis and thiols might have played in the emergence of biochemistry. AMP, adenosine monophosphate; R and R_1, amino acid side chain.



Biochemistry of nonribosomal peptide synthesis

Adenylation of an amino acid C terminus is followed by nucleophilic attack on the adenylated amino acid by the thiol (–SH) moiety of the phosphopantetheine group (red). Peptidyl carrier protein (PCP)—bound amino acid thioesters form peptide bonds, leading to peptide chain growth.



Catalytic aminonitrile-based peptide ligation

The –SH group of the cysteine catalyst forms a bond with an aminonitrile, (C≡N) to form an amino acid thioimidate intermediate. An amino acid then reacts with the intermediate, a peptidyl amidine bond is formed, and the catalyst is released.

GRAPHIC: A. KITTERMAN/*SCIENCE*

University of Strasbourg, Centre National de la Recherche Scientifique, Institut de Science et d'Ingénierie Supramoléculaires, 67000 Strasbourg, France. Email: moran@unistra.fr

Published by AAAS

cess quantities of sulfide (*13*). Such step-changes in reaction conditions pose a problem for a realistic early Earth scenario. However, by exploiting the fact that thiols reversibly add to α-amidonitriles, Foden *et al*. proposed a strategy for catalytic and redox-neutral oligopeptide synthesis proceeding via a thioimidate intermediate in neutral water (see the figure). The authors subjected acetylated glycinyl nitrile and free glycine to acetylated cysteine (a hypothetical phosphopantetheine precursor) at 60°C in neutral water. Within 24 hours, an acetylated Gly-Gly peptidyl amidine had formed, at a 60% yield, without the need for synthetic activating agents. This reactivity occurred with a variety of thiols and amino acids. Only amino-terminal serine, threonine, and cysteine could be coupled to α-amidonitriles stereoselectively, which might have implications for the origins of biological homochirality.

Foden *et al*. highlight the potential of small-molecule organocatalysis and the pivotal role that thiols might have played in the emergence of biochemistry. Could nature have made peptides this way before the advent of enzymes or the ribosome? This exciting question reveals how much we do not know about the rationale behind the shaping of life's biochemical pathways and the chemical structures used as cofactors and metabolites (*6*). If life did once use nitrile-based chemistry, it is unclear why it would have shifted to a biochemistry where nitriles are rare. Also, why does life—seemingly so wastefully—use the phosphopantetheine moiety in nonribosomal peptide synthesis, when it might have used a much simpler thiol, such as cysteine? Experimental insights into these questions will illuminate the currently opaque middle ground between reactions that attempt to recreate life's earliest chemistry and the biochemistry we study today. ■

REFERENCES AND NOTES

1. G. Danger, R. Plasson, R. Pascal, *Chem. Soc. Rev.* **41**, 5416 (2012).
2. C. S. Foden *et al*., *Science* **370**, 865 (2020).
3. M. Rodriguez-Garcia *et al*., *Nat. Commun.* **6**, 8385 (2015).
4. J. G. Forsythe *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7652 (2017).
5. S. A. Harrison, N. Lane, *Nat. Commun.* **9**, 5176 (2018).
6. J. E. McMurry, T. P. Begley, *The Organic Chemistry of Biological Pathways* (Roberts and Company Publishers, Inc., ed. 2, 2016).
7. A. Sauerwald *et al*., *Science* **307**, 1969 (2005).
8. K. B. Muchowska, S. J. Varma, J. Moran, *Chem. Rev.* **120**, 7708 (2020).
9. B. N. Khare, C. Sagan, *Nature* **232**, 577 (1971).
10. I. Shalayel *et al*., *Eur. J. Org. Chem.* 2020, 3019 (2020).
11. P. E. Dawson, T. W. Muir, I. Clark-Lewis, S. B. Kent, *Science* **266**, 776 (1994).
12. A. L. Weber, *Orig. Life Evol. Biosph.* **35**, 421 (2005).
13. P. Canavelli, S. Islam, M. W. Powner, *Nature* **571**, 546 (2019).

10.1126/science.abf1698

ROBOTICS

How to improve robotic touch

Challenges for instrumenting robotic hands to match human performance are outlined

By **Subramanian Sundaram**[1,2]

Human hands are densely covered with touch receptors called mechanoreceptors that provide us the sense of touch. This continuous tactile feedback from objects we touch, and information on hand articulation and movement—called proprioception—enable us to effortlessly handle diverse objects with fine dexterity. Bell's prescient treatise on the hand, written in 1833, refers to the human hand as the "consummation of all perfection as an instrument" (*1*). This sentiment echoes stronger today, in part fueled by the steep challenge of effectively instrumenting a robotic hand to provide similar feedback with the goal of attaining human-level dexterity. Recent insights into the primate tactile system and advances in machine learning (ML) may offer new prospects for tackling this old robotics challenge.

Roboticists have looked at all aspects of human tactile physiology (ranging from mechanoreceptors to neural coding schemes to grasping strategies) for inspiration on designing robotic touch systems. Most of the attention has focused on creating electronic sensors that mimic mechanoreceptors (see the figure) (*2*). Four types of mechanoreceptors—each responding to different types of tactile stimuli—innervate the skin: the fast adapting receptors FA I and FA II, and slowly adapting receptors SA I and SA II (*3*). The spatial coverage of these distinct mechanoreceptor types in the human hand have been systematically mapped since 1970, bringing attention to their collective functional roles. However, the subcellular molecular transducers of force have been harder to identify. The mechanically activated ion channels PIEZO1 and PIEZO2 were found to be force transducers in mammalian cells (*4*), and evidence of PIEZO2's role in human tactile and proprioceptive function is enabling an understanding of the initial signals involved in human touch (*5*). PIEZO2 is known to be involved in many low-threshold mechanoreceptors.

Currently, there are many unknowns in our understanding of the exact links between these initial transducers of force (ion channels) and mechanoreceptors. Many families of mechanosensitive ion channels other than PIEZOs are believed to contribute to touch because PIEZOs alone do not account for all human mechanosensation (*6*), but what these other families are, and how signals from these transducers are processed together, are not known. Even if these transducing channels were identified, it is not clear what role they play in enabling the distinct mechanoreceptor types that respond to specific stimuli. Mechanoreceptor responses may encode signals from many diverse force transducers, and a better understanding of the mammalian system may guide artificial sensor design.

First-order tactile neurons that collect signals from mechanoreceptors have a branched architecture, thereby integrating signals from multiple mechanoreceptors in distant points of the skin known as the receptive field (*3*). This complex spatial arrangement allows each first-order neuron to record object features during contact—like edges—as opposed to detecting forces at single points (*7*). As a result, the signal carried by a first-order neuron captures many properties as seen through multiple transducers and from many different locations. Furthermore, this signal may depend on material properties such as hardness and surface texture. Prior to sensing by mechanoreceptors, object contact is mediated by the skin, which modulates its own mechanical properties through skin hydration. Moisture secreted from sweat glands hydrates keratin in the skin and softens it. This process changes the contact dynamics between the fingertip and an object; junctional contact area increases more slowly when touching a hard, smooth glass surface as compared to a soft rubber (*8*).

Replicating all of the subtle features of mechanoreceptors and the skin in contemporary robots has remained daunting. However, substantial progress is being made in the creation of soft, stretchable electronic skins that can sense loads normal to the surface or shear forces against the surface; in both cases, static or dynamic forces can be sensed (*2*, *9*). Creative new fabrication strategies that use soft materials and high-performance organic electronics have also enabled neuromorphic sensors that produce action potential–like voltage spikes in response to forces (*10*). Neuromorphic tactile systems designed

[1]Biological Design Center, Boston University, Boston, MA, USA. [2]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA, USA. Email: subras@csail.mit.edu

Published by AAAS

for robots today are still nascent when compared with architectures in humans. In humans, the spatial distribution of receptors is typically 240 units/cm^2 at fingertips, with over 17,000 receptors covering the hand (*3*). Contemporary robots lag by two to three orders of magnitude, and the number of levels of distributed processing need much improvement. However, this approach is promising because spike-based processing can efficiently encode temporal relationships, motion, and other object properties with high fidelity and energy efficiency.

Passive touch is not the only way in which we acquire tactile information. To identify a grasped object, and simultaneously estimate its material characteristics, texture, and weight—all without the aid of vision—humans often use active haptic sensing, the use of exploratory and informative actions, to refine the sensory information acquired (*11*). Thus, we can determine texture by sliding our fingers, and reliably distinguish fabrics like velvet from wool.

Although robotic manipulation of objects with similar sensorimotor control is limited by tactile hardware in robot hands, there has been exciting progress in using visual data. The ubiquity of cameras and the ability to acquire large datasets—which are critical for effectively utilizing deep convolutional neural networks (CNNs) and reinforcement learning algorithms—are contributing to the success of visuomotor policies in manipulation tasks (*12*, *13*). Ultimately, relying too heavily on vision alone has its problems. This approach works well in fixed settings without any visual occlusions, such as robotic pick and place. Achieving manual dexterity in robot manipulation with vision alone may prove to be difficult in highly unstructured and dynamic settings without continuous line-of-sight access, such as searching through piles of rubble in disaster response.

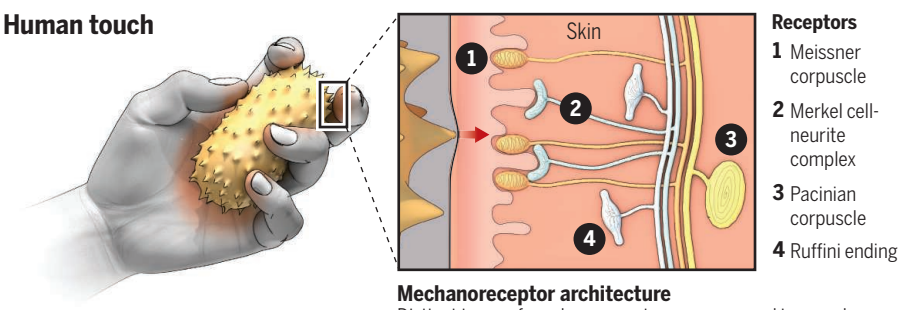
The success of robot vision–based object manipulation has many strong implications for robotic touch. It confirms that the emerging ML tools are effective at distilling information from high-dimensional images into actionable control policies. Many classes of algorithms such as CNNs can be readily used with tactile data (*14*), and vision-based ML strategies can be used for grasp planning.

A roadmap for hardware advances in tactile sensors may focus on at least three broad themes. There should be renewed emphasis on the reliability of tactile sensors over a robot's lifetime (several years), and on producing large amounts of high-quality data. Multimodal transducers that record light touch, deformation, vibrations, and temperatures using array architectures with increased distributed processing and reduced wiring are needed. Robot hands specifically suitable for sensorimotor control will benefit from a codesign of actuators and sensors with a high density of coverage (>100 sensors/cm^2 over large areas). In addition, robust proprioception will be needed in such applications for reliably localizing object contact.

Neuromorphic tactile hardware (and software) advances will strongly influence the future of bionic prostheses—a compelling application of robotic hands. Adhering to encoding techniques that closely mirror natural tactile coding schemes has been critical in providing users with realistic sensory feedback through peripheral nerve interfaces (*15*). In that regard, increasing the density of these connections will be crucial. ■

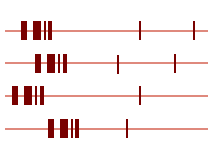
Human versus robotic touch

Human hands are covered with sensors (mechanoreceptors) that provide continuous touch-sensation feedback that allow us to identify and control objects. For example, we can exert the right amount of force to hold sharp and pointy fruits such as a horned melon (kiwano) with great poise and dexterity. Robotic hands do not have these capabilities but continue to improve.



Mechanoreceptor architecture
Distinct types of mechanoreceptors are arranged in complex architectures in the skin and record specific types of touch stimuli. Sweat can change the elasticity of skin and affect these signals.

Signal processing
Encoded tactile signals contain rich information about object contact, timing, and movements and collectively reflect object properties.

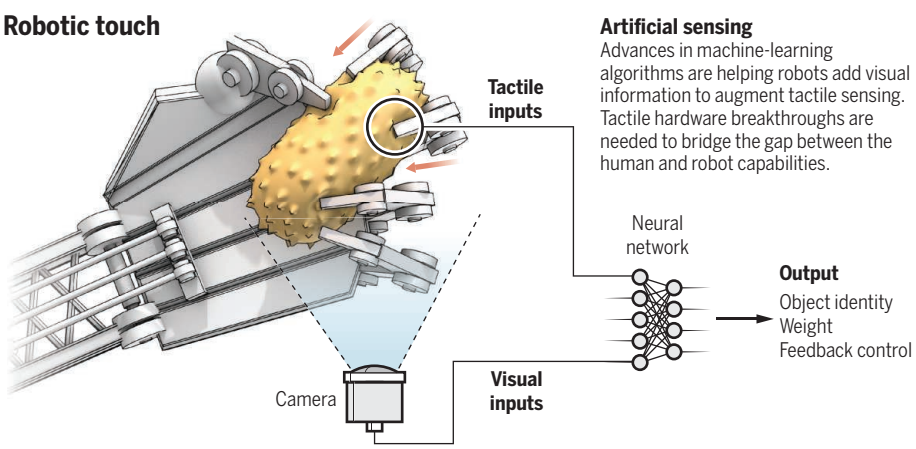


Touch signals encode timing and object properties.

Neural processing

Tactile signals and proprioception

Distilled information
Object identity
Weight
Feedback control



Artificial sensing
Advances in machine-learning algorithms are helping robots add visual information to augment tactile sensing. Tactile hardware breakthroughs are needed to bridge the gap between the human and robot capabilities.

GRAPHIC: C. BICKEL/*SCIENCE*

REFERENCES AND NOTES

1. C. Bell, *The Hand: Its Mechanism and Vital Endowments as Evincing Design* (The Bridgewater Treatises, William Pickering, London, UK, 1833), vol. 4.
2. C. Bartolozzi, L. Natale, F. Nori, G. Metta, *Nat. Mater.* **15**, 921 (2016).
3. A. B. Vallbo, R. S. Johansson, *Hum. Neurobiol.* **3**, 3 (1984).
4. B. Coste *et al.*, *Science* **330**, 55 (2010).
5. A. T. Chester *et al.*, *N. Engl. J. Med.* **375**, 1355 (2016).
6. F. Moehring, P. Halder, R. P. Seal, C. L. Stucky, *Neuron* **100**, 2 (2018).
7. J. A. Pruszynski, R. S. Johansson, *Nat. Neurosci.* **17**, 1404 (2014).
8. B. Dzidek, S. Bochereau, S. A. Johnson, V. Hayward, M. J. Adams, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 10864 (2017).
9. A. Chortos, J. Liu, Z. Bao, *Nat. Mater.* **15**, 937 (2016).
10. Y. Kim *et al.*, *Science* **360**, 998 (2018).
11. S. J. Lederman, R. L. Klatzky, *Cognit. Psychol.* **19**, 342 (1987).
12. S. Levine *et al.*, *J. Mach. Learn. Res.* **17**, 1334 (2016).
13. M. Andrychowicz *et al.*, *Int. J. Robot. Res.* **39**, 3 (2020).
14. S. Sundaram *et al.*, *Nature* **569**, 698 (2019).
15. J. A. George *et al.*, *Sci. Robot.* **4**, 32 (2019).

ACKNOWLEDGMENTS

S.S. is supported by American Heart Association grant 20POST35210045.

10.1126/science.abd3643

Published by AAAS

RETROSPECTIVE

James S. Jackson (1944–2020)

Social psychologist who pioneered the study of Black culture

By **Thomas A. Parham**

James S. Jackson, beloved elder and a light for Black culture in the field of social psychology, joined the community of Ancestors on 1 September at the age of 76. Jackson understood the need to explore the within-group variability of Black people's lives and culture rather than relying on simplistic between-group comparisons. His study of Black communities contributed a host of insights into family composition, education, health status and outcomes, aging, violence in the community, religious and spirituality practices, help-giving and help-seeking behaviors, law enforcement and policing, and experiences with racism. He reframed the discourse on people of African descent, giving Black people a voice on Black mental health.

Born in Detroit, Michigan, on 30 July 1944, Jackson completed his bachelor's degree in psychology at Michigan State University in 1966, his master's degree in psychology at the University of Toledo in 1970, and his Ph.D. in social psychology at Wayne State University in 1972. He then began his long career at the University of Michigan, from which he retired last year as a professor of psychology, professor of Afro-American and African studies, and director and research professor at the university's Institute for Social Research. His career included several stints on governmental advisory boards, including a position on the advisory council to the director of the National Institutes of Health.

In an effort to shed light on a group that was mischaracterized in the research and theories being presented at the time, Jackson established the Program for Research on Black Americans in 1976. The following year, he began his groundbreaking National Survey of Black Americans—a dataset that was arguably the largest and most comprehensive study of Black life and culture, as well as behavior patterns, in the country. Jackson's research methodology helped usher in a more authentic and realistic picture of the Black experience, and his data samples provided the platform for countless investigatory endeavors. His subject samples were not restricted to a hundred students enrolled in an introductory psychology course. Rather, his initiatives penetrated communities across the country, yielding population samples that few could match for their inclusive, decidedly Black diversity. He also challenged the implicit and sometimes explicit bias in university courses and journal editorial review boards that Euro-American norms were the standard against which Black people should be measured and compared. His efforts to create a more self-determined body of research on African American people, independent of their white counterparts, revolutionized research endeavors and data reporting in the 1980s and beyond. His scholarship also helped entire disciplines and a host of researchers understand that there are always multiple factors that explain a given phenomenon.

For Jackson, advocacy on behalf of Black Americans extended beyond research. As a graduate student, he was a member of the Black Student Psychological Association, which sought to create a cultural comfort zone for students. Later, he was part of a contingent of Black psychologists who confronted the American Psychological Association about its lack of membership diversity, the standard theoretical models in which Black people were invisible, and the testing and assessment protocols that inaccurately and inappropriately claimed to assess the intellectual capability and psychological stability of African American children and adults.

I considered Jackson my respected elder, an mzee, as those of us of African descent refer to those whom we hold in high esteem. I first encountered him on the pages of my textbooks. Any student of psychology with an interest in persons of African descent knew that Jackson's work was a must-read. His perspectives were interesting and insightful, his assertions compelling, and his theories comprehensive. I later met him at national meetings of the Association of Black Psychologists and the American Psychological Association. While his presence in the latter was an annual treat, it was his stature in the former that helped me truly understand the scope and impact of his work.

I always appreciated Jackson's orientation to an African-centered view of reality. Like many Black psychologists of his era, he embraced a strengths-based view of Black culture. This perspective was important, as oftentimes Black life viewed through a Eurocentric lens was assumed to be deficient or pathological because it was different from what white people thought or how they behaved, according to their own psychological experts. And yet, the salience of Jackson's work, in changing the focus on people of African descent from telephoto to wide-angle, allowed people throughout the country to realize and embrace a more holistic view of Black people's humanity.

Jackson had a quiet demeanor, a sharp intellect, and keen insight. He was tough-minded when it came to rigorous research questions but tender and empathic when revealing and contextualizing the core of Black people's life experience. Although he was averse to self-promotion, his work spoke for itself through the impact it had on how Black people perceived themselves and were perceived by others.

Jackson's awards and honors include his election as a fellow of the American Association for the Advancement of Science (AAAS, the publisher of *Science*), the American Academy of Arts and Sciences, and the New York Academy of Medicine. He was recognized for outstanding contributions to research in aging by the Gerontological Society of America's Robert W. Kleemeier Award. In 2014, he was appointed to the National Science Board.

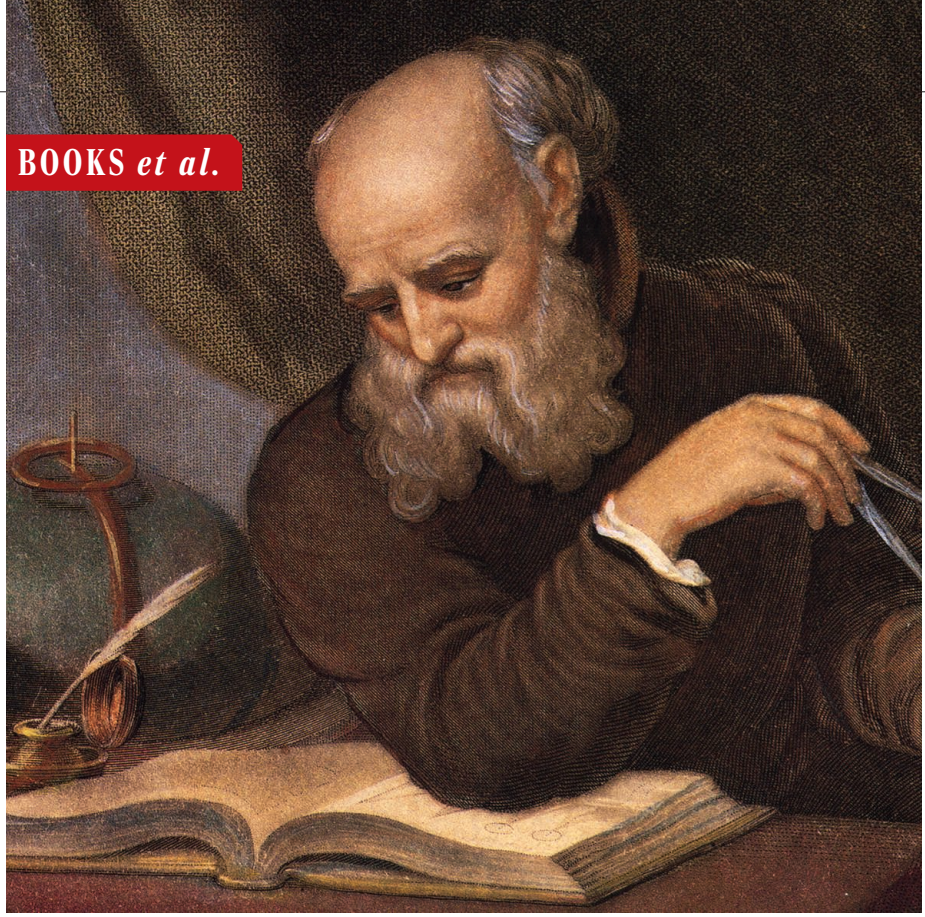
Just as the Sun never truly sets, Jackson's smile, ideas, and spirit will never truly leave us. His identity, character, and humanity have been indelibly imprinted on those places he traveled during his life, and his contributions have been permanently recorded in the pages of numerous manuscripts, books, scholarly journals, advisory boards, scientific meetings, and conference presentations. His collective works have served as an intellectual solvent, erasing profound misunderstandings and biases about people of African descent and revealing a rich and previously unappreciated cultural tapestry. ■

PHOTO: UNIVERSITY OF MICHIGAN INSTITUTE FOR SOCIAL RESEARCH

California State University Dominguez Hills, Carson, CA 90747, USA. Email: presidentparham@csudh.edu

10.1126/science.abf0706

Published by AAAS



Galileo's paradigm-shattering theories went beyond what is permitted by the scientific method.

IMAGE: PICTORIAL PRESS LTD/ALAMY STOCK PHOTO

CLASSICS REVISITED

Moving beyond the paradigm

A provocative 1975 call to leave behind dogmatic thinking could help scientists strengthen ties with the public

By **Joseph Swift**

Amid the turbulence of the 20th century's civil rights movement and sexual revolution, the philosophy of science was undergoing its own radical transformation. Suspecting that the scientific method was less straightforward than scientists claimed, philosophers had started challenging the idea that deductive logic was the best way to reveal truths about the world. *Against Method* (1975), by philosopher of science Paul Feyerabend, played a key role in bringing such arguments to maturity. Forty-five years after its publication, the book continues to offer valuable insights to scientists confused by the public's ambivalence toward hard scientific truths.

Most modern scientists would agree that the scientific method provides the best route toward an ever more cohesive understanding of the world around us. Feyerabend did not think it was so simple. He believed that within the landscape of all discoverable knowledge, the scientific method offers a path leading to only a fraction of all knowable facts. This is because it encourages researchers to begin where well-established theories leave off, keeping them aligned with existing scientific paradigms. A path set forward by classical physics, he argued, will not lead to quantum mechanics.

Against Method
Paul Feyerabend
New Left Books, 1975.
339 pp.

Feyerabend thought that new scientific paradigms could only be reached by radical methods. Indeed, given that new paradigms sit outside existing knowledge structures, there cannot be a predefined method on how to discover them. Anarchistic thinking, spiritualism, irrationality—all must remain on the table.

As proof of principle, Feyerabend elegantly demonstrated that a strict adherence to the scientific method would have forced Galileo to give up his hypothesis that Earth orbits the Sun. Not only did the existing evidence support the idea that Earth was stationary, the practice of science in the 17th century was largely entrusted to human perception. This meant that not being able to feel Earth moving would have been considered by many to be sufficient to falsify Galileo's theory. Feyerabend asserted that Galileo needed to break the existing scientific paradigm by presenting a new one and that he only succeeded in doing so by going beyond what rational argument allowed, drawing upon, for example, ad hoc hypotheses and emotional language.

Against Method was divisive. After publication, Feyerabend was called the "worst enemy of science" in *Nature* (*1*) and a "breath of fresh air" in *Science* (*2*). Many scientists thought that the book presented a type of philosophy that could be easily weaponized (*3*), not least because it provided a shield for nonexperts promoting unsubstantiated or malicious arguments. Some also worried that by weakening the boundaries of what counts as bona fide scientific research, science itself might come to be considered just another type of cultural practice.

But this is not quite what Feyerabend sought to inspire. Instead, he wanted to generate curiosity about what happens when we try to live within the rules of our current scientific paradigms. Is it always desirable, for example, to treat mathematical harmonies and statistical abstractions as the best reflections of reality? Consider empirical research on the state of American democracy, which largely relies on random sampling and quantitative metrics. By discounting narratives of police injustice as anecdotal, some have argued that political scientists long remained blind to the extent of racial authoritarianism (*4*).

Against Method hints that there must be a middle ground between one extreme, in which all views are equally valued, and the other, in which the limits of current scientific paradigms are never tested. This casts the role of today's scientist as more ambiguous than perhaps many would like, but such ambiguity could help scientists strengthen their relationship with the public. By loosening the framework within which scientists may operate, Feyerabend gives them permission to enter the political arena, a realm that many researchers have historically deemed outside their jurisdiction, but one in which their participation is sorely needed. ■

The reviewer is at the Salk Institute for Biological Studies, La Jolla, CA 92037, USA. Email: jswift@salk.edu

REFERENCES AND NOTES

1. T. Theocharis, M. Psimopoulos, *Nature* **329**, 595 (1987).
2. W. J. Broad, *Science* **206**, 534 (1979).
3. M. Kuntz, *EMBO Rep.* **13**, 885 (2012).
4. V. M. Weaver, G. Prowse, *Science* **369**, 1176 (2020).

10.1126/science.abe9322

Published by AAAS

HISTORY OF SCIENCE

Imaginary demons and scientific discoveries

A new history acts as a field guide to the fictional imps invoked by scientific thinkers

By **Jan G. Michel**

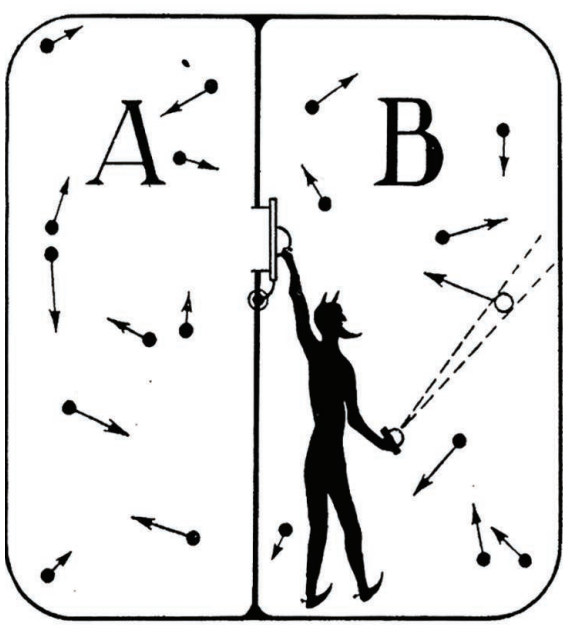
What role could imaginary creatures possibly play in the most rational, systematic, and discovery-oriented enterprise humanity has ever created? Did the Enlightenment not teach us to dispense with the supernatural and provide us with the power of reason to exorcise demons from modern science?

As Jimena Canales reveals in her thought-provoking and highly readable book *Bedeviled*, fictional imps that help us explore the limits of what is possible pervade the history of scientific thinking, namely in the form of thought experimentation. Although this is not a new topic of inquiry, Canales comes at it from a refreshingly original perspective, examining the ways that various scientific thinkers throughout the centuries have made use of imaginary demons.

In the book's 10 chapters, Canales recounts science's most famous fiends, including Laplace's demon, a being who could predict every occurrence in the Universe, and Maxwell's demon, who could violate the second law of thermodynamics. More contemporary examples are discussed in the book's later chapters and include the self-programming electronic circuits that Norbert Wiener referred to as "demons" and the lines of code known as "daemons" that are central to today's computer communication infrastructure.

At the outset of her scientific demonology, Canales presents a classic example: Descartes's demon. In his *Meditations on First Philosophy* (1641), the French philosopher, mathematician, and scientist René Descartes attempted to establish a firm foundation for scientific knowledge. To do so, he used a method known as "Cartesian doubt," which encouraged a skeptical approach to reality. After arguing that our sensory experience is often erroneous and also that we can never be certain that we are not dreaming, Descartes imagined an even worse scenario: What if a powerful and clever demon systematically presented us with a false reality? If this were the case, we could not trust our senses at all. There is, however, one feature of reality that such a demon would be unable to manipulate. "Cogito, ergo sum," reasoned Descartes, "I think, therefore I am." From this foundation, Descartes constructed an approach to scientific knowledge that fundamentally influenced Western philosophy and modern science.

In her chapter on computer daemons, Canales discusses Searle's demon, an imp introduced by philosopher John Haugeland in response to an assertion made by philosopher John Searle that computers do not understand the symbols they manipulate (*1*). Searle argued that artificial intelligence is comparable to a native English speaker with no knowledge of the Chinese language, who is locked in a room and asked to provide sensible responses in Chinese to inquiries made in that language. Given an English rule book that instructed the user how to correlate Chinese symbols with appropriate responses, the person could convincingly convey that he understood the messages being relayed without actually understanding them. Similarly, argued Searle, the central processor of a computer translates without understanding any of the symbols it manipulates. But, asked Haugeland, what happens if "a superfast person"—Searle's demon—took over an individual's neural processing, replicating that person's exact responses? What is then causally responsible for the person's understanding—her neurons or the demon?



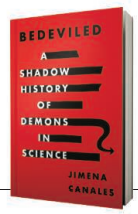
Maxwell imagined a "very observant and neat fingered" demon who could detect and react to the motions of individual molecules.

We learn a great deal about the scientists who have used demons to confront difficult challenges and approach apparent paradoxes. For instance, along with her description of the article in which Albert Einstein invoked "ghosts" in order to question traditional notions of space (and, later, time), Canales includes details about Einstein's health problems, his broken marriage, and how Arthur Eddington helped to confirm his theory of general relativity.

The reader also learns about the role that cultural factors played in shaping scientific thought at different times and in different contexts. For example, we learn why Descartes warned against the "dangers" of reading novels like those that intrigued Don Quixote and also about the role played by the Renaissance painter Raphael in Erwin Schrödinger's conceptions of life and negative entropy.

As Canales maintains, demons are "neither just psychological delusions or simplistic heuristic fictions nor simply auxiliary midwives who help scientists deliver knowledge." On the contrary, she argues that the notion of a (scientific) demon should be regarded as akin to other, more established philosophical notions or tools, "such as concepts, numbers, classes, and categories"—a welcome contribution to the philosophy of scientific discovery that deserves further scholarly attention. ■

Bedeviled: A Shadow History of Demons in Science
Jimena Canales
Princeton University Press, 2020. 416 pp.



The reviewer is at the Department of Philosophy of Religion and Science, Ruhr-Universität Bochum, 44801 Bochum, Germany. Email: jagumi@gmx.de

REFERENCES AND NOTES

1. J. R. Searle, *Behav. Brain Sci.* **3**, 417 (1980).

10.1126/science.abd9851

IMAGE: REPRINTED BY PERMISSION FROM SPRINGER NATURE

Published by AAAS

Disease transmission can occur at live animal markets, but zoonotic disease research could benefit from an emphasis on humans' and animals' shared risk of infection.

PHOTO: AMILIA ROSO/THE SYDNEY MORNING HERALD/GETTY IMAGES

Edited by **Jennifer Sills**

EcoHealth reframing of disease monitoring

Decade-old (*1*) and recent warnings for coronaviruses with zoonotic epidemic potential (*2*) could have prevented the emergence of coronavirus disease 2019 (COVID-19) (*3*). We therefore agree with Watsa and colleagues ("Rigorous wildlife disease surveillance," Perspective, 10 July, p. 145) that wildlife biosurveillance should increase. However, representing animals as a threat to humans through disease transmission leads to ill-conceived reactive policies (*4*). A perspective (*5*) in which animals and humans share similar risks of pathogens and infections, making animals relevant disease models and sentinels, would be more effective. Clarifying the connection between animal and human health could increase public support for research seeking to understand host-switching in animals, such as the study of virus evolution (*6*), interactions in pathogen communities (*7*), and pathogen discovery (*8*).

A shared-risk perspective on emerging infectious diseases mirrors the field of EcoHealth, which explores the links between ecosystem, animal, and human health. Such strategies place value in healthy ecosystems through an integrative approach that considers both pathogen biodiversity and social-ecological drivers (*9*). Prevention based on understanding the transmission of pathogens through EcoHealth-based emerging infectious disease surveillance is a promising avenue for sustainability science, orders of magnitude cheaper than mitigation in response to a transfer to human hosts (*10*), and less intrusive than current crisis responses.

Maarten P.M. Vanhove[1,2,3]*, Jean Hugé[4,5,6], Luc Janssens de Bisthoven[7], Hans Keune[8,9], Anne Laudisoit[10], Séverine Thys[11], Erik Verheyen[12,13], Nicolas Antoine-Moussiaux[14]

[1]Research Group Zoology: Biodiversity and Toxicology, Centre for Environmental Sciences, Hasselt University, Diepenbeek, Belgium. [2]Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic. [3]Laboratory of Biodiversity and Evolutionary Genomics, Department of Biology, University of Leuven, Leuven, Belgium. [4]Department of Environmental Science, Open University of the Netherlands, Heerlen, Netherlands. [5]Department of Biology, Vrije Universiteit Brussel, Brussels, Belgium. [6]Research Group Environmental Biology, Centre for Environmental Sciences, Hasselt University, Diepenbeek, Belgium. [7]Capacities for Biodiversity and Sustainable Development, Royal Belgian Institute for Natural Sciences, CEBioS program, Brussels, Belgium. [8]Belgian Biodiversity Platform–Research Institute Nature and Forest, Brussels, Belgium. [9]Chair Care and the Natural Living Environment, Department of Primary and Interdisciplinary Care Antwerp, Faculty of Medical and Health Sciences, University of Antwerp, Antwerp, Belgium. [10]Ecohealth Alliance, New York, NY 10018, USA. [11]Department of Vaccinology, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium. [12]Operational Directorate Taxonomy and Phylogeny, Royal Belgian Institute for Natural Sciences, Brussels, Belgium. [13]University of Antwerp, Department of Biology, Evolutionary Ecology, Antwerp, Belgium. [14]Fundamental and Applied Research for Animals and Health, Faculty of Veterinary Medicine, University of Liège, 4000 Liège, Belgium.
*Corresponding author.
Email: maarten.vanhove@uhasselt.be

REFERENCES AND NOTES

1. V. C. C. Cheng, S. K. P. Lau, P. C. Y. Woo, K. Yung Yuen, *Clin. Microbiol. Rev.* **20**, 660 (2007).
2. Y. Fan *et al.*, *Viruses* **11**, 210 (2019).
3. European Environment Agency (EEA), "Late lessons from early warnings: Science, precaution, innovation—Summary" (Report 1/2013, Publications Office of the European Union, Luxembourg, 2013); www.eea.europa.eu/publications/late-lessons-2.
4. N. Antoine-Moussiaux *et al.*, *Sustain. Sci.* **14**, 1729 (2019).
5. P. M. Rabinowitz, L. Odofin, F. J. Dein, *EcoHealth* **5**, 224 (2008).
6. S. J. Anthony *et al.*, *Virus Evol.* **3**, vex012 (2017).
7. W. de Souza, *Parasitol. Res.* **119**, 2369 (2020).
8. D. R. Brooks *et al.*, *WCSA Journal* **1**, 1 (2020).
9. H. Lerner, C. Berg, *Front. Vet. Sci.* **4**, 163 (2017).
10. A. P. Dobson *et al.*, *Science* **369**, 379 (2020).

10.1126/science.abe8239

Build international biorepository capacity

In their Perspective "Rigorous wildlife disease surveillance" (10 July, p. 145), M. Watsa *et al.* underscore the value of One Health approaches to stimulate integration

Published by AAAS

across currently siloed efforts in zoonotic research and mitigation. To achieve comprehensive decentralized pathogen surveillance, there is an urgent need to develop environmental and biodiversity infrastructure in biodiverse countries experiencing high rates of habitat conversion, wildlife trafficking, and human-wildlife interactions.

Approximately one-third of One Health networks lack an environmental component, fewer than half are active in wildlife surveillance, and almost none is led by developing countries (*1*). International support for development of natural history museums with frozen vertebrate tissue collections remains a key component missing from the One Health equation. Most pathogens causing severe outbreaks in humans are zoonotic in origin (*2*); thus, understanding their evolution and that of their wild animal hosts is imperative.

As was the case for coronavirus disease 2019 (COVID-19) (*3*), identifying wild animal reservoirs can be challenging when biorepositories are lacking (*4*). In most countries, natural history biorepositories remain poorly supported and largely disconnected from public health initiatives. For example, most studies of bat coronaviruses to date (*5*), including the PREDICT animal surveys discussed in Watsa *et al.*, did not preserve host specimens or tissues, thus limiting the potential for molecular host identification or replication and extension of the science (*6*). Emerging infectious disease response hinges on sampling depth across space, time, and taxonomy, the very sampling enabled by museum biorepositories. As primary biological infrastructure, in-country development of museum collections that follow best practices (*7*), with specimen data freely available through the internet, should be an international imperative (*8*) for effective global surveillance and mitigation of emerging infectious diseases.

Jocelyn P. Colella[1], Bernard Risky Agwanda[2], Faisal Ali Anwarali Khan[3], John Bates[4,5], Carlos A. Carrión Bonilla[6,7], Noé U. de la Sancha[4,8], Jonathan L. Dunnum[7], Adam W. Ferguson[4], Stephen E. Greiman[9], Prince Kaleme Kiswele[10], Enrique P. Lessa[11], Pamela Soltis[12], Cody W. Thompson[13], Maarten P. M. Vanhove[14,15,16], Paul W. Webala[17], Marcelo Weksler[18], Joseph A. Cook[7]*

[1]Biodiversity Institute, University of Kansas, Lawrence, KS 66045 USA. [2]National Museums of Kenya, Nairobi, Kenya. [3]Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia. [4]Field Museum, Chicago, IL 60605, USA. [5]Natural Science Collections Alliance, Washington, DC 20005, USA. [6]Museo de Zoologiá, Escuela de Biología, Pontificia Universidad Catolica del Ecuador, Quito, Ecuador. [7]Museum of Southwestern Biology and Biology Department, University of New Mexico, Albuquerque, NM 87131, USA. [8]Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA. [9]Department of Biology, Georgia Southern University, Statesboro, GA 30458, USA. [10]Centre de Recherche en Sciences Naturelles, Lwiro, Democratic Republic of Congo. [11]Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay. [12]Florida Museum of Natural History and the University of Florida Biodiversity Institute, University of Florida, Gainesville, FL 32611, USA. [13]Department of Ecology and Evolutionary Biology and the Museum of Zoology, University of Michigan, Ann Arbor, MI 48108, USA. [14]Research Group Zoology: Biodiversity and Toxicology, Centre for Environmental Sciences, Hasselt University, Diepenbeek, Belgium. [15]Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic. [16]Laboratory of Biodiversity and Evolutionary Genomics, Department of Biology, University of Leuven, Leuven, Belgium. [17]Department of Forestry and Wildlife Management, Maasai Mara University, Narok, Kenya. [18]Departamento de Vertebrados, Museu Nacional, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.
*Corresponding author. Email: cookjose@unm.edu

REFERENCES AND NOTES

1. M. S. Khan *et al.*, *Lancet Planet. Health* **2**, e264 (2018).
2. K. E. Jones *et al.*, *Nature* **451**, 990 (2008).
3. J. Cohen, *Science* 10.1126/science.abd7707 (2020).
4. S. A. J. Leendertz, J. F. Gogarten, A. Düx, S. Calvignac-Spencer, F. H. Leendertz. *EcoHealth* **13**, 18 (2016).
5. B. Hu *et al.*, *PLoS Path.* **13**, e1006698 (2017).
6. J. A. Cook *et al.*, *Bioscience* **70**, 531 (2020).
7. J. L. Dunnum *et al.*, *PLoS Negl. Trop. Dis.* **11**, 1 (2017).
8. O. Paknia, H. Sh Rajaei, A. Koch, *Organ. Divers. Evol.* **15**, 619 (2015).

10.1126/science.abe4813

Response

We agree with Vanhove *et al.* that wildlife conservation and emerging infectious disease screening are two sides of the same coin. Wildlife and humans can be vulnerable to spillover events by the same pathogen. For example, respiratory diseases (*1*) and Ebola virus (*2*) outbreaks have occurred simultaneously in great apes and humans. Pathogens also affect biogeographical species range expansions, contractions, and extinctions (*3*). Biosurveillance efforts should reflect that health risks are shared by humans and wildlife, a central tenet of the One Health framework (*4*). As Vanhove *et al.* point out, wildlife can serve as the source for preventive solutions that mitigate spillover risks into humans and animals.

A shared risk perspective could also combat the narratives that portray animals as dangerous pests or disposable commodities that endanger human health (*5*), as in the case of bats (*6*), many of which are likely not hosts for coronaviruses such as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (*7*). In addition to emphasizing shared risk, such misinformation can be countered with well-researched messaging following a zoonotic outbreak. Conservation social science has honed a suite of tools to identify the often unpredictable human motivations behind (*8*), and the possible negative consequences of, such communications (*9*).

Colella *et al.* suggest that surveillance efforts should include natural history collections. Some natural history museums and zoos archive biobanked specimens, cryopreserved viable cell cultures, disease specimen banks, and histopathology samples, but this highly effective practice (*10*) is limited by high costs. We agree that devoting funding toward biodiversity banking within countries at high risk for emerging infectious diseases would improve conservation outcomes. Taxonomically diverse biobanked tissues and live cell cultures could expand studies of host-pathogen relationships, clarifying host range or affected tissues and providing in vitro systems for infectivity and pathogenicity investigations. Such collections could allow drug development for humans to expand beyond just a few animal laboratory models, given that relatively well-studied viruses such as SARS-CoV-2 are potentially broadly infectious across taxonomic orders (*11*). Comparative genomics and transcriptomics among nonmodel species are used infrequently in biomedical research programs but hold great potential for prioritizing species and gene targets with alternative host defense mechanisms for laboratory study (*12*).

Mrinalini Watsa[1,2]* and Wildlife Disease Surveillance Focus Group[3]

[1]Population Sustainability, San Diego Zoo Global, San Diego, CA 92027, USA. [2]Field Projects International, San Diego, CA 92126, USA. [3]Wildlife Disease Surveillance Focus Group authors and affiliations are listed at science.sciencemag.org/content/369/6500/145/suppl/DC1.
*Corresponding author.
Email: merkenswickwatsa@sandiegozoo.org

REFERENCES AND NOTES

1. J. D. Negrey *et al.*, *Emerg. Microbes Infect.* **8**, 139 (2019).
2. S. A. J. Leendertz *et al.*, *Mamm. Rev.* **47**, 98 (2017).
3. R. E. Ricklefs, E. Bermingham, *Glob. Ecol. Biogeogr.* **11**, 353 (2002).
4. P. M. Rabinowitz *et al.*, *Ecohealth* **5**, 224 (2008).
5. J. P. Kibambe *et al.*, "In Africa, wildlife raises the risk of deadly diseases: It doesn't have to" *CNN* (2020).
6. H. Zhao, *Science* **367**, 1436 (2020).
7. H. Yan *et al.*, biorxiv 10.1101/2020.09.08.284737 (2020).
8. K. E. Wallen, E. Daut, *Nat. Conserv.* **26**, 55 (2018).
9. H. N. Dang Vu, M. R. Nielsen, *Hum. Dimensions Wildl.* **23**, 417 (2020).
10. J. Radin, *J. Cult. Econ.* **8**, 361 (2015).
11. J. Damas *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 22311 (2020).
12. L. Yurkovetskiy *et al.*, *Cell* **183**, 739 (2020).

10.1126/science.abe9392

Published by AAAS

COOLING IN A WARMING WORLD

NEWS
Living with heat p. 778

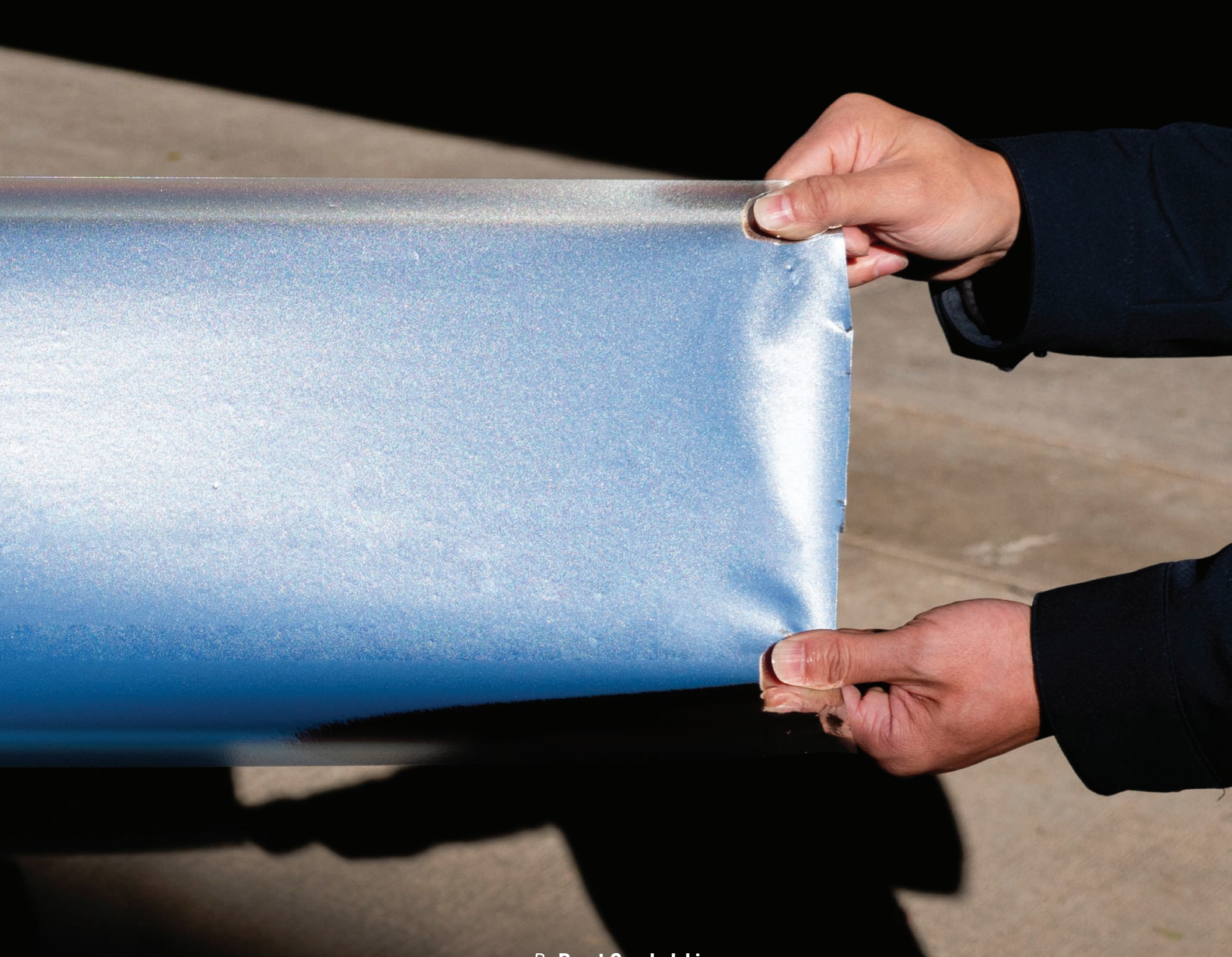
PERSPECTIVES
Adapting to the challenges of warming p. 782
Cooling our insatiable demand for data p. 783
Photon-engineered radiative cooling textiles p. 784

REVIEWS
Terrestrial radiative cooling: Using the cold universe as a renewable and sustainable energy source p. 786
New refrigerants and system configurations for vapor-compression refrigeration p. 791
Caloric materials for cooling and heating p. 797

RELATED ITEMS Podcast

PHOTO: GLENN ASAKAWA/UNIVERSITY OF COLORADO

Published by AAAS



By **Brent Grocholski**

The planet is warming from greenhouse gas emissions, with increases in both peak and average temperatures. Consequently, the demand for cooling will intensify, driven by climate change as well as improving economic conditions. Cooling requirements for data centers and new technologies may also call for innovative solutions. Developing a wide array of cooling strategies will be vital for human comfort and prevention of heat-induced medical emergencies. A growing number of materials and methods are being designed to improve technologies, meet the temperature needs of data centers and other facilities, and reduce harmful emissions.

Vapor compression systems supply much of the residential and industrial cooling infrastructure. These systems rely on compressing gases, many of which contribute to the greenhouse effect if released into the atmosphere. Finding replacements is a priority but requires balancing trade-offs between toxicity, flammability, and lower efficiency. An alternative approach involves the use of caloric materials—i.e., solids that effectively pump heat as they are electrically, magnetically, or mechanically manipulated.

A completely different cooling strategy is to use materials that regulate heat in the form of infrared radiation. Passive radiative cooling materials take advantage of an atmosphere window for infrared radiation to eject heat into space. When deployed on rooftops, these systems can potentially cool buildings by a few degrees, even in the daytime. Along the same lines, textiles may be modified so that heat can more efficiently radiate through them, allowing people to be more comfortable at higher thermostat set points.

As the planet warms, keeping cool without increasing greenhouse gas emissions will be a challenge, but new materials and technologies are poised to meet this goal.

A large-area film engineered to passively radiate heat away from the surface

Published by AAAS

LIVING WITH HEAT

Studies are probing how heat threatens health—and how to lower the risks

By **Elizabeth Pennisi**

It's 5 a.m. and still dark at the Carlton Complex fire camp in central Washington, except for the fire's orange glow on a distant ridgeline. Wildlands firefighter Bre Orcasitas, two colleagues, and three volunteers suit up: heavy duty fire-resistant pants, shirt, jacket, and helmet. Their boots weigh 2 kilograms; the backpacks they will haul to the fire—loaded with 6 liters of water, food for a 16-hour shift, safety gear, and hand tools—can weigh 30 kilograms. Sometimes the burden includes a 12-kilogram chain saw.

On this day in August 2014, the crew is not just fighting flames, but also taking part in research. Orcasitas outfits each person with a chest harness and sensors that will record their heart rate, elevation gain, distance traveled, carbon monoxide intake, and skin temperature. Each swallows an ingestible radio thermometer that relays deep body temperature to the chest monitor every 15 seconds via Bluetooth. Orcasitas and her two colleagues will record each firefighter's activities, be it cutting down trees, digging a fire break, or burning vegetation to keep a larger fire away. It's all part of a study to assess heat exposure in wildlands firefighters—the biggest ever to do so. From 2013 through 2016, more than 300 firefighters participated.

High body temperatures are inevitable in firefighting: A study in 2013 uncovered about 50 heat-related injuries across the United States during that fire season. But other data from their project have surprised Orcasitas and her colleagues. Warmth from the firefighters' physical exertion, not heat from the fires, was the greatest danger, the researchers found. Another surprise: "The assumption across the fire community was that if somebody went down, it was because they just didn't drink enough water," Orcasitas says. But the team found otherwise. "You can't drink yourself out of a heat-related injury," explains project leader Joseph Domitrovich, an exercise physiologist at the U.S. Forest Service's National Technology and Development Program. "It's not the magic bullet that people thought."

Across the globe, researchers like Domitrovich are working to pin down how heat affects workers and vulnerable populations, such as the elderly. They are studying low-tech measures—sometimes nothing more than a splash of cold water on the skin—to make people safer and more comfortable in hot conditions. And they are exploring the body's ability to adapt to the heat. "Deaths and illnesses caused by heat are largely preventable," says June Spector, an occupational and environmental health physician-scientist at the University of Washington, Seattle.

The work has taken on urgency as global temperatures rise, heat waves become more frequent and intense, and casualties mount. Between 1999 and 2010, the U.S. Centers for Disease Control and Prevention logged 8081 heat-related deaths in the United States, one-third of them in people age 65 or older. Already, about one-third of the world's population experiences conditions that create heat stress, says Nathan Bradley Morris, a human thermal

PHOTO: PAUL JONES/UNIVERSITY OF WOLLONGONG

Published by AAAS

Sweat plays an important role in cooling the body, so researchers are capturing and measuring it.

physiologist at the University of Copenhagen. At high risk are construction workers and farm hands, respectively 13 times and 35 times as likely to die from heat as other workers. Farm hands in the United States already face an average of 21 days in summer when humidity and temperature combine to exceed recommended limits, and the risks are growing. In the United States, climate change will double the number of unsafe days by 2050 and triple it by 2100, Spector and colleagues reported on 25 August in *Environmental Research Letters*.

Amplifying the trends, the world's population is moving to cities, which tend to be hotter than the countryside, says Lisa Leon, a research physiologist at the U.S. Army Research Institute of Environmental Medicine. In Paris, about 12,000 people died in 1 week during a heat wave in 2003. Even those who don't work in the heat are vulnerable, especially people who are older and overweight or whose hearts may not be strong enough to cope with the physiological stress of heat, says Christopher Gordon, a retired physiologist from the Environmental Protection Agency. And much of the global population is aging and getting heavier.

By 2100, heat exposure will threaten the health of an estimated 4 billion people, according to an International Labour Organization report. The report forecasts that the world's productive working hours will decline 2.2% because of increasing heat, resulting in $2.4 trillion in economic losses, concentrated in Southern Asia and Western Africa. "There are inequities in who is exposed to heat," Spector says.

FOR CENTURIES, physiologists have studied how the human body reacts to heat stress, tracing its remarkable ability to keep its core temperature at about 37°C even when the outside air is many degrees hotter. Sensory nerves on the skin respond to the environment's temperature, and internal sensors keep tabs on heat gain, some of which is generated by the body itself. Intense physical activity can turn the body into a furnace, raising heat production 15-fold. Whether the source is environmental or internal, if heat is not dissipated, it can strain and ultimately shut the body down.

When internal sensors tell the brain the body is warming, the hypothalamus sends signals that dilate blood vessels close to the skin, causing more blood to circulate there and lose heat—provided the air is cooler than the body. When it's not, or if heat transfer to the air isn't enough, the sweat glands get to work. Another part of the brain, the medulla oblongata, gets in touch with the heart, which increases its rate and the amount of blood pumped per beat. The body's fluids shift, redirecting blood to the skin and helping fill sweat glands.

An active person can easily sweat out 2 liters of water per hour, which cools the body as it evaporates. It's not the head, as often believed, but the hands (with their high concentration of sweat glands) and the torso (with its large surface area) that are key sites for sweat-aided cooling. The fluid loss stimulates the release of hormones that enhance thirst and alter kidney function to reduce urine production. These fluid-conserving measures help shore up blood volume so the heart can maintain blood pressure at safe levels.

Dehydration greatly adds to heat strain and the risk of serious injury. Staying well-hydrated can be protective, but it may not keep body temperature from rising if internal or external heating is too great, Domitrovich and others have found. First comes heat exhaustion, characterized by cramps, fatigue, headache, nausea, or dizziness—signals that prompt most people to move out of the heat or stop exercising. As long as sweating continues, the body can still cool itself somewhat. But if a person gets too hot (this tipping point varies, but hovers around 42°C for exertional heat stress), sweating can stop and body temperature can skyrocket, sometimes even rising above 44°C. The brain falters; confusion, agitation, slurred speech, even coma can result.

Even if people recover, heat can cause problems. Some 15% of people exposed to chronic heat stress at work develop kidney problems, possibly making it one of the first epidemics due to global warming, a European consortium called Heat Shield reported in 2018 in *The Lancet Planetary Health*. Moreover, studies have documented an increase in other injuries on hot days, possibly because heat and dehydration can impair thinking.

At their worst, life-threatening temperatures can damage the brain and cause organs to fail. Deprived of normal blood flow, the gut can leak, causing widespread inflammation. Blood vessels can get damaged and blood can coagulate. Cells can even fall apart as their proteins break down.

"Heat stroke is one of the three leading causes of death for athletes, soldiers in training, and laborers," says Douglas Casa, an exercise physiologist at the University of Connecticut (UConn), Storrs.

That's what happened to Korey Stringer, an offensive lineman for the Minnesota Vikings. He collapsed on day two of spring training in 2001, but his heat stroke was not treated aggressively, and he died the next day. A professional athlete, he was fit—but not fit enough for such a strenuous workout early in the season. Motivation to exercise may override the body's signals to stop, Leon says.

Gatorade and the National Football League launched the Korey Stringer Institute at UConn in 2010, and now grants and private support keep it going. As its CEO, Casa has worked extensively with athletes and with the U.S. military to protect recruits from Stringer's fate. For the most part, the military and sports trainers have "a good handle on treating heat stroke, but I don't think people are doing enough to prevent it from happening in the first place," Casa says.

At the institute, suited up with heavy backpacks, rifles, and full military attire, volunteers march around a lab warmed to the temperature of a hot desert. Casa studies how bodies react to the stress of working in the heat and tests protective measures, including new fabrics for hot weather clothing, new safety and recovery procedures, and wearable sensors that can sound the alarm if conditions reach dangerous extremes.

Published by AAAS

In a study last year, for example, some of the volunteers got thirsty during a bout of fast walking on a treadmill in a room at 35°C with 30% humidity. Afterward, experimenters allowed the subjects to drink enough to replace 25% of their lost fluids and asked whether they were still thirsty. They were not. Thirst was quenched well before they had enough water, Casa and his colleagues reported in November 2019 in *Nutrients*. "The absence of thirst does not mean the absence of dehydration," he warns. He recommends that elite athletes and others who exert themselves in hot conditions figure out their sweat rate and adjust their drinking accordingly.

Long slogs and heavy clothing and gear can put firefighters at even greater risk of heat injury than the fire itself.

Casa has also field tested a strategy for runners suffering from exertional heat stroke. For decades, his team has provided medical care for a 12-kilometer race on Cape Cod in August. The race is short enough that runners keep up a punishing pace but long enough that their bodies can overheat; whereas about one in 10,000 marathon racers develops heat stroke, far more—one in about 650—racers in this event do so. At the Cape Cod race, his team treats up to 45 heat stroke victims a year by immersing them in cold water. Getting their body temperature below 40°C within 30 minutes is enough to ensure a full recovery, the team has found. "Most people would be surprised to know we can treat heat stroke," Casa says. In fact, he says, the treatment guarantees recovery if used promptly and properly.

FAR BETTER to avoid heat strain entirely, says Nigel Taylor, a thermal physiologist retired from the University of Wollongong. That means adjusting to heat by avoiding air conditioning, which Taylor says "prevents us from adapting to our climate."

Heat tolerance varies from person to person, not just because of age and health, but also because of genetic factors. One study of 42,000 Indigenous miners in South Africa when they were first sent to work in the hot mines found that about 15% could not handle the heat, whereas 25% coped just fine.

But Taylor and others have found that frequent exposure to heat also makes a big difference to heat tolerance. If an individual has a chance to get used to being hot—by spending time in hot weather or exercising strenuously—the body modifies its physiology and becomes less vulnerable to heat stress. Athletes and workers can then work harder and longer despite the heat.

He and others have studied adaptation in detail. They find that after just 1 week spending 2 hours a day working outdoors in a hot environment, the body begins to adjust. Normal deep body temperature drops. The body sweats at a lower temperature, and there is less strain on the heart. That's because, even in the heat, heart rate doesn't rise as much, and the heart pumps more blood per stroke. The body retains more fluids and blood volume rises, increasing water reserves for sweating and cooling. "The body has a really good capacity to acclimate," says Daniel Gagnon, a human physiologist at the University of Montreal. But "as long as you keep inserting air conditioning in that process, you delay acclimatization," says Elizabeth Repasky, an immunologist at Roswell Park Comprehensive Cancer Center.

In the heat of summer, there may be other ways to cool down, says Ollie Jay, a thermal physiologist at the University of Sydney. He has put octogenarians, children, people with heart conditions, even pregnant women into a heat chamber and has found that fans can often be as effective as air conditioning, at much lower environmental and financial costs. Jay and his team reported in November 2019 in the *Annals of Internal Medicine* that fans can be effective at up to 40°C, particularly in humid conditions, where they help evaporate sweat that would otherwise sit on the skin. (In dry environments, sweat evaporates regardless of fan use.)

Workers can be as productive and comfortable when cooled by fans instead of air conditioners, he and his colleagues reported last year in *Energy and Buildings*. "In terms of increases in work productivity, fan use in a simulated Vietnamese working environment of 30°C with 70% humidity is the equivalent of 7°C of cooling with air conditioning," Jay says.

Splashing cold water on the skin can also work well, his studies show. "External dousing does the job of sweat without having to sweat," which can lead to dehydration and strain the heart, Jay says. His team evaluated dousing and fans as alternatives to air conditioning during this summer's U.S. heat wave. In 80 of 105 cities—the exceptions being in the Southwest—those alternative measures would have been 100% effective, he and his colleagues reported on 25 July in *Science of the Total Environment*. "It's getting hotter, and heat waves are getting worse," he says. "What's important is using these findings to make changes in public policy."

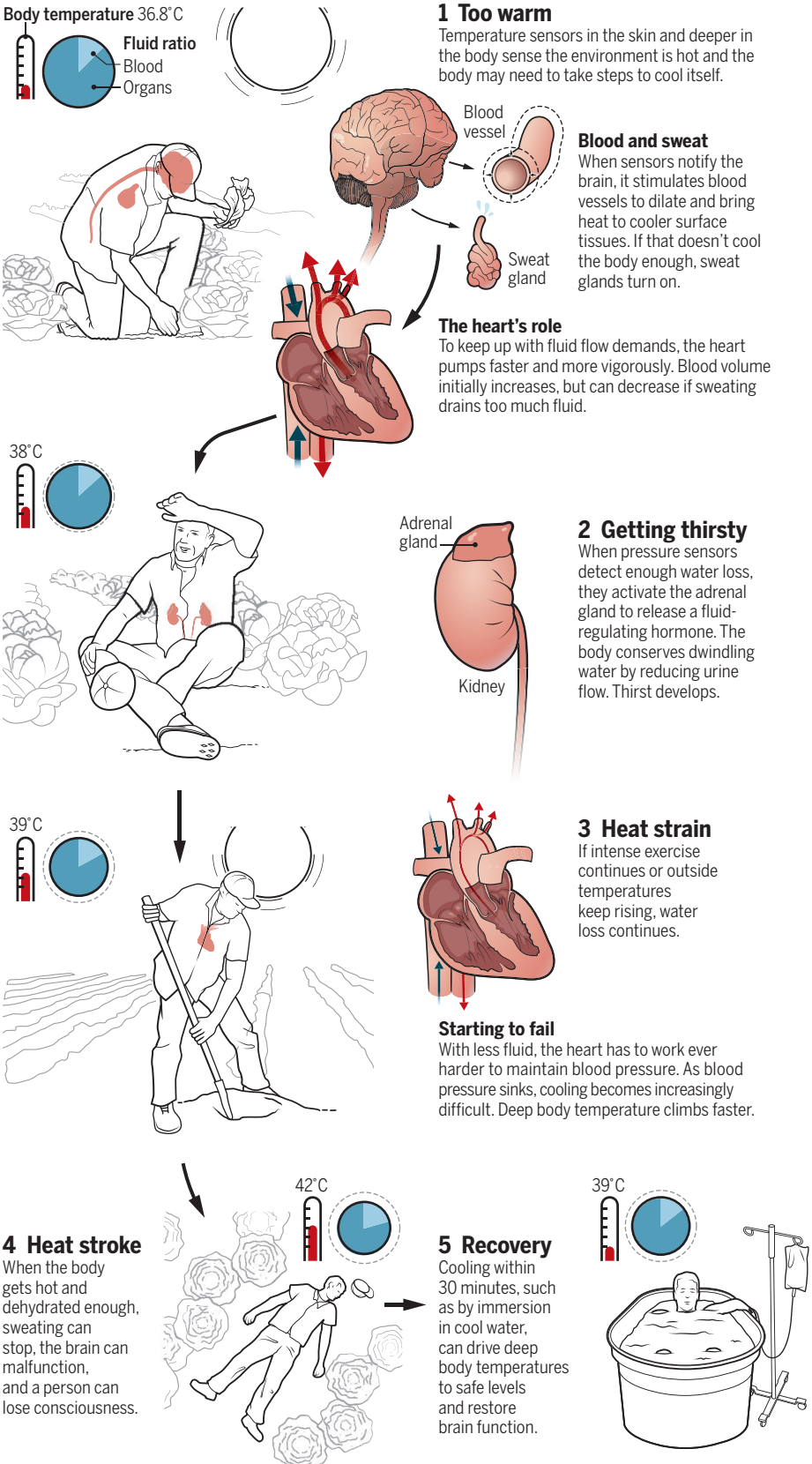
For outdoor workers, one simple—but not always easy—fix is to move to the shade. In a rural part of Indonesia, Spector,

PHOTO: DAVID MCNEW/GETTY IMAGES

Published by AAAS

When cooling fails

The body works hard to keep its temperature stable. If deep body temperature rises, organs work together to drive heat to the skin. As heating continues, the heart, skin, brain, and other organs try to maintain the right body temperature and amount of fluids. But those systems can be overwhelmed. (Normal body temperature and the danger threshold vary from person to person.)


Body temperature 36.8°C
Fluid ratio
Blood
Organs
1 Too warm
Temperature sensors in the skin and deeper in the body sense the environment is hot and the body may need to take steps to cool itself.
Blood vessel
Sweat gland
Blood and sweat
When sensors notify the brain, it stimulates blood vessels to dilate and bring heat to cooler surface tissues. If that doesn't cool the body enough, sweat glands turn on.
The heart's role
To keep up with fluid flow demands, the heart pumps faster and more vigorously. Blood volume initially increases, but can decrease if sweating drains too much fluid.
38°C
Adrenal gland
Kidney
2 Getting thirsty
When pressure sensors detect enough water loss, they activate the adrenal gland to release a fluid-regulating hormone. The body conserves dwindling water by reducing urine flow. Thirst develops.
39°C
3 Heat strain
If intense exercise continues or outside temperatures keep rising, water loss continues.
Starting to fail
With less fluid, the heart has to work ever harder to maintain blood pressure. As blood pressure sinks, cooling becomes increasingly difficult. Deep body temperature climbs faster.
42°C
4 Heat stroke
When the body gets hot and dehydrated enough, sweating can stop, the brain can malfunction, and a person can lose consciousness.
5 Recovery
Cooling within 30 minutes, such as by immersion in cool water, can drive deep body temperatures to safe levels and restore brain function.
39°C

GRAPHIC: V. ALTOUNIAN/*SCIENCE*

Yuta Masuda and Nicholas Wolff of the Nature Conservancy, and colleagues randomly assigned 363 laborers to work in a forest or a deforested area nearby. (Deforestation can increase local temperatures by up to 8°C.) The workers wore heart rate monitors and had their oral temperatures taken regularly to calculate core body temperature. During a 90-minute task in hot, humid, sunny conditions, people in open areas had temperatures above 38.5°C for 3 minutes longer than those in the forested area, Spector and her colleagues reported last year in *Environmental Research Letters*. The differences may seem small, but over the course of a day, body temperatures would likely keep climbing, putting these workers at great risk of heat-related illness. Those who worked in deforested areas also scored worse on cognition and memory tests, perhaps because of dehydration or discomfort, the team reports in a paper in press in *Environmental Research Letters*.

Wildlands firefighters also work in an exposed, hot environment. Yet it is not sunshine or fire they need to worry about most. "The heat from the fire almost never has any effect on our core body temperatures," says William Knudsen, a wildlands firefighter based in Helena-Lewis and Clark National Forest. Nor is it dehydration, once considered the biggest risk for these workers. Orcasitas found, for example, that after a particularly strenuous day's work, a well-hydrated firefighter came back feeling "like crap" and had a higher body temperature than a colleague who started and ended the day more dehydrated. And fellow study coordinator Knudsen saw firsthand how trouble can arise even in temperate weather. His team recorded its highest body temperatures on a high-elevation site with an air temperature of only about 21°C.

Instead, Domitrovich says, "It's the internal heat production that's most critical." Big packs, heavy clothing, and a fast pace cause body temperatures to spike. "And if you are less fit, you are going to generate more heat," Orcasitas says. So now, when Knudsen takes out his crew, he's mindful of the weight they carry and how fast they hike, frequently taking a 2-minute break after 8 minutes of walking.

Orcasitas thinks the study will make a difference for her and her firefighting colleagues. "These are all seemingly simple, but [the findings] are a pretty big deal," she says, because they're easy to act on.

They could also help others stay cooler as global temperatures climb. "Heat stress is a very complex topic, so the more we can educate not just firefighters but everyone, the better," Domitrovich says. "It will lead to increased safety for all humans." ■

Published by AAAS

PERSPECTIVE

Adapting to the challenges of warming

The impact of regional heat events is becoming more important to quantify

By **Steven C. Sherwood**

Heat extremes on Earth have reached a disturbing new level in recent years. The July 2020 temperatures soared across Siberia and reached a record-breaking 38°C inside the Arctic Circle, continuing a line of record heat events globally. "Event attribution" calculations, which are an endeavor to apportion blame for extreme events through quantitative modeling, suggest that some events would have been nearly impossible without human-induced global warming. This includes the recent Siberian summer and the 2018 heat wave in Japan, which killed more than a thousand people (*1*, *2*). Rising heat is creating new challenges for humanity that will require new adaptation and protection measures. Smart implementation requires careful calculation of how further global temperature rises will translate into short-term regional heat events and how these will translate into impacts on human health and activities, food supply, infrastructure, and ecosystems.

These enhanced heat extremes are the result of slightly more than 1°C of global-mean anthropogenic warming since the mid-19th century. The world has set a goal under the Paris Agreement to keep global-mean anthropogenic warming below 2°C in the future, or 1.5°C if possible. Current national commitments, however, could well permit increases of up to 4°C or so by the end of this century, with further increases thereafter (*3*). Even the ambitious Representative Concentration Pathway (RCP) 4.5 scenario considered by the Intergovernmental Panel on Climate Change in 2013 leaves only a 50% chance of remaining below 2.4°C by 2090, given the latest estimates of climate sensitivity (*4*). Limiting carbon dioxide to meet the 2°C target will therefore require some combination of substantially stricter emissions-reduction commitments and policies or carbon dioxide removal efforts, each presenting daunting sociopolitical, economic, and technological challenges.

A common saying is that people do not feel the average temperature. However, climate model projections do indicate that in most regions, peak temperatures (see the figure) roughly track the annual mean in the same location (*5*). In some regions such as Western Europe, extremes are predicted to increase faster than the mean owing to greater variability. Moreover, the mean increases vary around the globe; one reason is that the upwelling of cold water from the deep oceans will cause Southern Hemisphere temperatures to lag behind those in the Northern Hemisphere for decades or centuries. Also, land regions generally warm more than oceans, a result of the difference in background humidity (*6*).

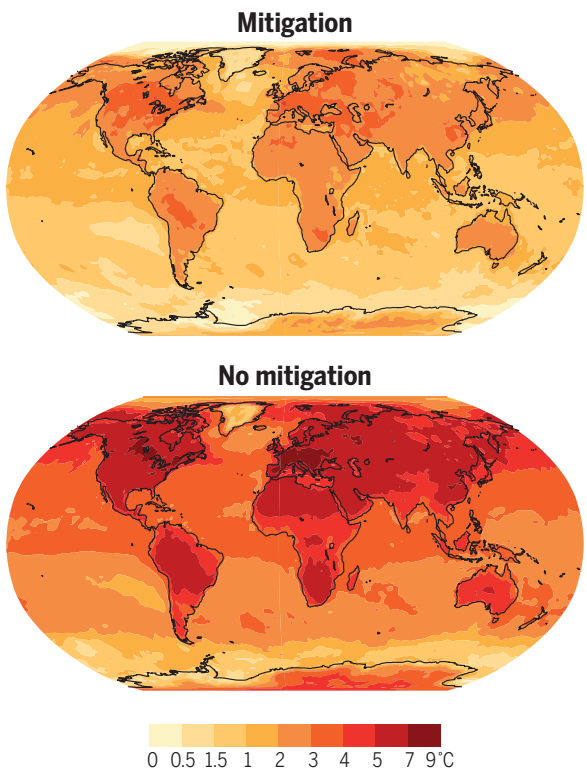
Although early research focused on maximum temperatures, additional factors have come to the fore. Humidity is crucial to heat stress, for example, because it inhibits evaporative cooling. Humidity is increasing globally along with temperature. Regional variations in these variables tend to compensate, so that areas like Western Europe and South America that become drier also warm more (*7*). Thus, heat stress will increase relatively more uniformly and predictably, whereas moisture stresses will change more variably, as some regions warm and dry substantially. Animals, including humans, and plants can be subject to either stress independently. Anthropogenic warming has roughly tripled the size of the human population that experiences dangerous humid heat annually (*8*), and this heat is already approaching the limits of human tolerance in a few locations (*9*). Humidity increases not only thermal stress outdoors but also the overall energy requirements of cooling systems, which already account for a large fraction of total power consumption in humid climates where air conditioning has become widespread (*10*).

Another factor to consider is the duration of heat events. More frequent or longer heat events tend to have greater impacts on human health (*11*), and nighttime temperatures are more strongly correlated to heat-related mortality than daytime ones. These observations show that adequate physiological recovery from heat exposure is important. Wind systems, such as the mid-latitude jet streams and associated weather systems, are expected to slow in a warmer climate, potentially causing longer lasting periods of extreme weather. However, multiple factors are at work such that the net change remains unclear (*12*).

Because of the "urban heat island" effect, temperatures are higher in urban areas than in surrounding natural areas (*13*). The relative lack of vegetation means that incoming solar energy goes into heating surfaces rather than canopies aloft, and less goes into evaporating water vapor. Human-made surfaces

Future rises in peak temperature

The increase in the maximum 20-year return value of maximum daytime temperature late this century (2081–2100) relative to 1986–2005, based on the average of many climate models, is shown. Projections based on a strong mitigation scenario [Representative Concentration Pathway (RCP) 4.5] (top) and a high-emission scenario (RCP8.5) (bottom) are shown.



GRAPHIC: ADAPTED FROM COLLINS *ET AL.* (*3*)

Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, NSW 2052, Australia. Email: s.sherwood@unsw.edu.au

Published by AAAS

and urban canyons also retain heat better into the night. In addition, the energy used in cities generates heat, which is negligible on a global scale but often important in urban areas. Urban heat islands are not typically included in climate calculations and are likely to worsen anywhere that urbanizes further, adding to the warming delivered by the climate system.

Although predicting the above factors is challenging enough, quantifying their impacts is even harder. Quantitative models of heat-affected natural and human systems, if used at all, are less advanced relative to the complexity involved than are weather and climate models. Meanwhile, climate change is creating conditions that lie outside the range of past experiences, limiting the reliability of empirical studies. Current impact models diverge substantially in the predicted impacts of climate changes (*14*) and almost surely suffer from systematic biases (*15*). Diverse impacts generally depend on the different aspects of heat events, devaluing any one-size-fits-all heat measure. We need to more rigorously quantify the links between meteorological forecasts and practical consequences.

Past studies do point to a couple of robust conclusions. One is that impacts will increase nonlinearly with mean warming, as extreme thresholds are crossed with rapidly increasing frequency (*8*). This highlights the need for strong emissions mitigation to keep warming to a level that we can cope with. The other is that although no one will be spared, the world's poor will be hit particularly hard (*11*). This highlights the need for low-cost adaptations and technologies as we seek suitable countermeasures to rising heat. ■

REFERENCES AND NOTES

1. World Weather Attribution Project, Siberian heatwave of 2020 almost impossible without climate change (2020); www.worldweatherattribution.org/siberian-heatwave-of-2020-almost-impossible-without-climate-change/.
2. Y. Imada, M. Watanabe, H. Kawase, H. Shiogama, M. Arai, *Sci. Online Lett. Atmos.* **15A**, 8 (2019).
3. M. Collins *et al.*, in "Climate change 2013: The physical science basis," T. F. Stocker *et al.*, Eds. (Report, Intergovernmental Panel on Climate Change, Cambridge Univ. Press, 2013), chap. 12.
4. S. C. Sherwood *et al.*, *Rev. Geophys.* **58**, e2019RG000678 (2020).
5. A. Di Luca, R. de Elía, M. Bador, D. Argüeso, *Weather Clim. Extrem.* **28**, 100255 (2020).
6. M. P. Byrne, P. A. O'Gorman, *Proc. Natl. Acad. Sci. U.S.A.* **115**, 4863 (2018).
7. E. M. Fischer, R. Knutti, *Nat. Clim. Chang.* **3**, 126 (2013).
8. D. Li *et al.*, *Environ. Res. Lett.* **15**, 064003 (2020).
9. C. Raymond *et al.*, *Sci. Adv.* **6**, eaaw1838 (2020).
10. E. Morakinyo *et al.*, *Renew. Energy* **142**, 73 (2019).
11. A. Ahmadalipour, H. Moradkhani, M. Kumar, *Clim. Change* **152**, 569 (2019).
12. T. Shaw *et al.*, *Nat. Geosci.* **9**, 656 (2016).
13. K. Deilami, M. Kamruzzaman, Y. Liu, *Int. J. Appl. Earth Obs. Geoinf.* **67**, 30 (2018).
14. S. G. Yalew *et al.*, *Nat. Energy* **5**, 794 (2020).
15. J. Schewe *et al.*, *Nat. Commun.* **10**, 1005 (2019).

10.1126/science.abe4479

PERSPECTIVE

Cooling our insatiable demand for data

Hyperscale data centers provide challenges and opportunities for energy use

By **Amy S. Fleischer**

The world is driven by constant access to information. We have the entire world at our fingertips wherever we are, and access to data in its various forms is ubiquitous. The ability to retrieve this information, much of it now stored in the cloud, is supported by data centers of various sizes and scales. Although instant access to data from anywhere is a benefit in many ways, part of the cost of this access is in escalating energy demands, much of it from the cooling infrastructure needed to support these data centers.

The widespread movement to cloud-based services over the past 14 years has transformed the data center industry. In 2006, 50% of all U.S. servers were located in small or medium-size data centers or in self-managed corporate data centers (*1*). These small data centers had an average processor utilization of only 10 to 20% and required a substantial energy investment because of high power demands. Inefficient cooling infrastructure for these servers, often located in closets or small rooms, led to a total data center annual energy demand of 70 terawatt hours in the United States in 2014 or about 1.8% of all U.S. energy usage (*1*).

The total energy demand of data centers includes both the energy demands of the IT equipment itself and the energy demand of the supporting infrastructure, the vast majority of which is for the cooling systems. A metric known as power usage effectiveness (PUE) represents the ratio of the total data center facility energy use to the energy use of the IT equipment. A data center with a PUE of 1.0 would use no energy other than that used to power the IT equipment. In 2007, when the metric was introduced by the Green Grid, the average data center PUE was 2.0, and this average value had not changed appreciably by 2016 (*1*). However, an increased focus on this metric in recent years has led to many advances in cooling technology.

College of Engineering, California Polytechnic State University, San Luis Obispo, CA, USA. Email: afleisch@calpoly.edu

Notably, these improvements in cooling architecture are implemented mainly in hyperscale data centers and have coincided with a widespread shift away from small, local, on-site data centers. In 2010, 79% of all data center operations were executed in traditional data centers, including small on-site data centers, whereas in 2018, 89% of compute instances were executed in cloud computing facilities (*2*). The economy of scale in hyperscale data centers permits the use of advanced server designs and vastly improved cooling technologies, and PUE can now approach 1.1 to 1.2 (*1*). These massively scaled data centers have doubled in number in just the past 4 years, from just over 250 worldwide in 2015 to more than 540 today, with hundreds more in development (*3*).

Data center cooling technology has come a long way from the server rooms of yore. Traditional designs often relied on air cooling and the overprovision of room air conditioning, necessitating the use of a parka for technicians working in the space. Improvements in air-handling design thus were the first major steps forward in data center cooling. Raised floor designs with integrated air distribution plenums were implemented, with cold air distribution routed directly into the intake of the servers by using perforated floor tiles in front of each rack or cabinet.

Subsequent advances included arranging servers into "hot aisle"–"cold aisle" configurations to reduce the mixing of the exhaust from one server with the intake of another; detailed computational fluid dynamics analysis of the floor tile designs to optimize air flow and minimize pressure drop (*4*); and the implementation of advanced systems and component-level control strategies for air handling. The benefit of some of these techniques is that they are suitable for retrofit into existing legacy data centers (*5*) while also being relevant to new designs. Hyperscale data centers, which by nature have flexibility in their geographic location, can be located in cool and low-humidity climates in which both air-side economization and water-side economization can be used to reduce cooling costs. Air-side economization features the mixing of outdoor air into

Published by AAAS

the conditioned air and brings the same benefits as opening the windows when cool outside. As such, the strategy is often referred to as "free cooling." Water-side economization reduces the demand on the chillers that cool the conditioned air in many data centers, again by using outside air. Evaporative cooling techniques can reduce the energy demand even further. By using such techniques, Google has dropped its average trailing 12-month PUE to 1.10 (*6*).

Future improvements in cooling strategies will come from a shift in focus from air-cooling strategies to a focus on embedded liquid and evaporative cooling techniques, which take advantage of inherently higher heat transfer coefficients. Direct liquid cooling features microchannel cold plates mounted directly on the heat-producing chips. Reliability concerns related to the integration of liquids in the server chassis initially affected the implementation of this technology, but increases in chip energy density beyond the capabilities of air cooling, coupled with technology advances such as optimized microchannel designs (*7*) and detailed failure mode analysis and mitigation (*8*), have led to wider adoption of embedded liquid cooling in hyperscale data centers over the past few years.

However, advances in integrated chip designs with three-dimensional architecture are creating energy densities beyond even that which can be controlled with direct liquid cooling. Thus, next-generation cooling research is investigating techniques such as flow and pool boiling (*9*) and even direct immersion of the chips in dielectric fluids with subsequent boiling at the chip interface (*10*). Thermosyphons, which feature gravity-fed liquid-vapor systems, are the focus of much attention (*11*), as they allow the high heat transfer possible with flow boiling while eliminating the reliability issues inherent with micropumps. In application, microevaporators are located directly at each chip, much like a cold plate; gravity-fed liquid feeds the evaporator, and the resulting vapor rises to a condenser located at the top of the server rack, where the loop begins again. Much two-phase flow research focuses on the use of environmentally friendly refrigerants such as HFC-245fa or HCFO-1233zd(E), adding the benefit that any leak will immediately vaporize, eliminating concerns around liquid leakage (*12*).

As the demand for data center capacity continues unabated, the demand on the power grid will continue to escalate. Thus, solutions for environmentally friendly and off-grid power production are of much interest. Some hyperscale data center operators are exploring sourcing their own sustainable power production using hydrogen fuel cells (*13*), and others are sourcing their energy from solar and wind farms. Future research trends will include a focus on advanced cooling techniques that feature energy recovery and reuse. Both liquid and two-phase cooling strategies produce a stream of heated liquid or vapor that can be used in various waste-heat recovery systems, including straightforward techniques such as hot water production and plant or district heating (*14*) and more complex systems such as absorption refrigeration or the organic Rankine cycle, which generates a source of electricity that can be fed back into the data center (*15*).

"As the demand for data center capacity continues unabated, the demand on the power grid will continue to escalate."

The current technical limitation is the low quality of the heat, and next-generation advances will require much attention to the integration of the cooling technology and the energy generation technique to meet the growing demands of our insatiable need for data. The past 5 years have shown incredible advancements in the technology used to cool data center equipment, but much work remains to be done to develop sustainable, environmentally friendly solutions to the energy demand required to power the cloud. ■

REFERENCES AND NOTES

1. A. Shehabi *et al*., "United States Data Center Energy Usage Report" (Lawrence Berkeley National Labs, LBNL-1005775, 2016).
2. E. Masanet, A. Shehabi, N. Lei, S. Smith, J. Koomey, *Science* **367**, 984 (2020).
3. www.srgresearch.com/articles/hyperscale-data-center-count-reaches-541-mid-2020-another-176-pipeline; accessed 10 October 2020.
4. J. Athavale, Y. Joshi, M. Yoda, *J. Electron. Packag.* **140**, 010902 (2018).
5. R. Khalid, A. P. Wemhoff, *J. Electron. Packag.* **141**, 041004 (2019).
6. www.google.com/about/datacenters/efficiency/; accessed 15 October 2020.
7. B. Ramakrishnan, Y. Hadad, S. Alkharabsheh, P. R. Chiarot, B. Sammakia, *J. Electron. Packag.* **141**, 041005 (2019).
8. S. Alkharabsheh, U. L. N. Puvvadi, B. Ramakrishnan, K. Ghose, B. Sammakia, *J. Electron. Packag.* **140**, 020902 (2018).
9. H. Lee *et al*., *J. Electron. Packag.* **140**, 020906 (2018).
10. R. Eiland *et al*., *J. Electron. Packag.* **139**, 041005 (2017).
11. F. Cataldo, J. R. Thome, *J. Electron. Packag.* **140**, 021002 (2018).
12. S. Araya, A. P. Wemhoff, G. F. Jones, A. S. Fleischer, *Appl. Therm. Eng.* **180**, 115757 (2020).
13. G. Saur *et al*., *Technical Report NREL/TP-5400-75355* (2019).
14. K. Ebrahimi, G. F. Jones, A. S. Fleischer, *Renew. Sustain. Energy Rev.* **31**, 622 (2014).
15. S. Araya, A. P. Wemhoff, G. F. Jones, A. S. Fleischer, *J. Electron. Packag.* **143**, 021001 (2021).

10.1126/science.abe5318

PERSPECTIVE

Photon-engineered radiative cooling textiles

Personal thermal management offers a path to reduce climate control energy use

By **Po-Chun Hsu** *and* **Xiuqiang Li**

Take a few seconds to look around and list all the technologies that are indispensable for you. If your list does not include textiles, try to live a typical day without them. Textiles are arguably one of the earliest human inventions. Without textiles to cover the human body for warmth, our ancestors would not have been able to spread across the various climate zones of the Earth. Today, many textiles are made for social etiquette and aesthetic purposes, but the pressing threat of global warming has created demand for innovative textiles that help to better cool the person who wears them.

The rationale behind linking textiles and climate change is that wearing the cooler textiles for localized "personal thermal management" may reduce the demand for air conditioning. The impact of air conditioning is considerable, given that it is not only responsible for 10% of U.S. electricity consumption but that the refrigerants are also a source of high global-warming-potential gasses (*1*, *2*). Considering an indoor setting distinguishes the new generation of cooling textiles from textiles oriented for sports. On average, the metabolic heat rate of indoor light activities is 60 to 80 W/m^2, balanced by the heat flux from the skin to the environment. This flux occurs through all viable heat transfer pathways: conduction, convection, radiation, and evaporation. Because one of the criteria of thermal comfort is the absence of sensible perspiration, evaporation only accounts for ~5 W/m^2. Using the American Society of Heating, Refrigerating and Air-Conditioning

Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA. Email: pochun.hsu@duke.edu

Published by AAAS

Engineers Standard 55 as the reference and assuming the clothing insulation is 1 clo (0.155 m^2K/W; clo is the industry unit for thermal insulation), then the heat transfer through textile conduction and natural convection contributes ~40 W/m^2, and radiation is responsible for ~25 W/m^2. This calculation demonstrates the substantial role of radiation in the human body heat balance. Unlike convective heat transfer, heat radiation is a surface property and does not require any media or moving part, making it a perfect tool for personal thermal management.

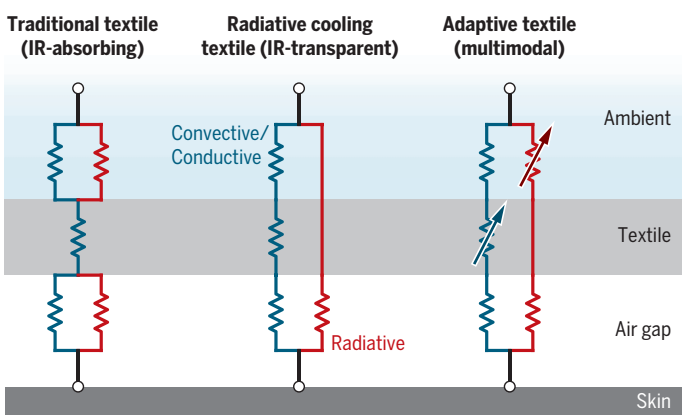
Regardless of skin pigmentation, the human skin is a nearly perfect black body that emits thermal radiation through Planck's law. This radiation transmits through the air gap and is absorbed by the textile, which contains various types of resonating molecular vibrational modes. The textile then reemits to the ambience. Almost all clothing materials are highly absorbing in the spectral region of human body radiation. Although Kirchhoff's radiation law indicates that they are also good emitters, their infrared (IR) opacity inevitably results in the radiation shielding effect. Therefore, rather than engineering the existing clothing materials, the key to radiative cooling is to re-invent the material so that it is transparent in mid-IR, allowing the thermal radiation from the hot human skin to bypass the textile and directly reach the ambience (see the figure). By contrast, traditional IR-absorbing textiles emit from the cold outer surface with a much lower radiation power that is proportional to the fourth power of temperature, according to Stefan-Boltzmann's law. This design principle leads to polyethylene (PE) textiles because of its simple chemical bonds and very few resonance peaks in mid-IR. In 2015, Tong *et al.* numerically predicted that radiative cooling fabric could be achieved by controlling the PE fiber diameter and light-scattering mechanisms to obtain both IR transparency and visible light opacity for wearability (*3*). One year later, Hsu *et al.* experimentally demonstrated nanoporous PE (nanoPE) radiative cooling fabric (*4*). The nanoporous structure was designed to achieve multifunctionality such as visible opacity, breathability, and softness. Several efforts to enhance the wearability were also conducted, including sweat-wicking and wind permeability. Built on the intrinsic IR transmittance and nanoscale photonic engineering, this concept was further developed into large-scale woven and knitted textiles (*5*), outdoor cooling textiles (*6*), and visibly colored cooling textiles (*7*, *8*). Depending on the types, these radiative cooling textiles can have the cooling performance equivalent to more than a 2°C increase of indoor temperature setpoint. One should not underestimate this amount of setpoint increase because it applies to the entire building space, which has orders of magnitude more thermal inertia and heat loss to manage compared with the occupants. Estimates of energy savings suggests the potential for a 20% reduction of heating, ventilation, and air conditioning energy with this small setpoint increase (*9*).

Looking ahead from the promise of energy saving, challenges and opportunities exist for the radiative cooling textiles. Thousands of years of evolution have made the expectation for clothing design to be complicated and often subjective in terms of what is considered necessary. Even for simple energy-efficient thermal comfort, psychological and environmental factors are in play that go beyond what might be ideal from a physical science perspective. The choices of IR-transparent polymers should be broadened to accommodate the various needs for wearability, such as moisture transport, skin touch comfort, scratch durability, and laundering. A general rule is to search for materials with a high content of crystalline aliphatic segments, followed by quantitative measurement of absolute extinction coefficient. Also, textiles should be adaptive. For example, the textile woven from carbon nanotube–coated bimorph yarns can respond to sweat and modulate the mid-IR emissivity (*10*), which is energy free and effective for the outdoors, but needs to increase the trigger sensitivity for the indoor scenario. A different strategy is to use active cooling by devices such as Peltier coolers that are either directly wearable or connected with recirculating water (*11*, *12*). These systems have superior cooling power, but the power consumption likely needs an energy storage or supply breakthrough to become part of daily clothing. A hybrid solution is variable passive thermoregulation, which uses energy to control the heat transfer coefficients rather than supply the thermal power. The tunable range is similar to the purely passive approach but can be actively changed according to user preference and potentially other signal inputs from the environment or the human body (*13*, *14*). Ultimately, the radiative thermal engineering can be combined with other textile heat-management mechanisms to accomplish multimodal control. A radiative cooling textile with coupled dynamic evaporative cooling can perform human body cooling even when the ambient temperature is higher than the skin.

Like many other renewable energy technologies, radiative cooling textile was initially a materials science and nanophotonics effort for niche applications for energy-efficient buildings by accomplishing the localized personal cooling, and it will eventually need to find the proper market position to be economically sustainable. In particular, these textiles will be joined by both traditional textile engineering and the booming wearable technologies. Our ancestors invented textiles as the "secondary skin" for thermal regulation. As textiles become an indispensable part of our lives, they hopefully will move closer to being as smart, versatile, and natural as our actual skin. ■

Designing more comfortable textiles

Heat transport through traditional textiles occurs by conduction or convection (blue paths), but infrared (IR) radiation is blocked (red paths). Cooling textiles improve radiative heat transfer between the skin and the environment. Adaptive textiles control the heat balance without working fluids or continuous energy input.



GRAPHIC: N. CARY/*SCIENCE*

REFERENCES AND NOTES

1. U.S. Department of Energy, "ARPA-E DELTA Program Overview" (2013); https://arpa-e.energy.gov/sites/default/files/documents/files/DELTA_ProgramOverview.pdf.
2. U.S. Energy Information Administration, "Annual Energy Outlook 2020" (2020) https://www.eia.gov/outlooks/aeo/.
3. J. K. Tong *et al.*, *ACS Photonics* **2**, 769 (2015).
4. P.-C. Hsu *et al.*, *Science* **353**, 1019 (2016).
5. Y. Peng *et al.*, *Nat. Sustain.* **1**, 105 (2018).
6. L. Cai *et al.*, *Adv. Mater.* **30**, 1802152 (2018).
7. L. Cai *et al.*, *Joule* **3**, 1478 (2019).
8. L. M. Lozano *et al.*, *Opt. Mater. Express* **9**, 1990 (2019).
9. T. Hoyt, E. Arens, H. Zhang, *Build. Environ.* **88**, 89 (2015).
10. X. A. Zhang *et al.*, *Science* **363**, 619 (2019).
11. S. Hong *et al.*, *Sci. Adv.* **5**, eaaw0536 (2019).
12. L. Lou *et al.*, *Energy Build.* **226**, 110374 (2020).
13. P.-C. Hsu *et al.*, *Sci. Adv.* **3**, e1700895 (2017).
14. E. M. Leung *et al.*, *Nat. Commun.* **10**, 1947 (2019).

10.1126/science.abe4476

Published by AAAS

REVIEW

Terrestrial radiative cooling: Using the cold universe as a renewable and sustainable energy source

Xiaobo Yin[1,2]*, Ronggui Yang[3]*, Gang Tan[4], Shanhui Fan[5]

Photonic materials designed at wavelength scales have enabled a range of emerging energy technologies, from solid-state lighting to efficient photovoltaics that have transformed global energy landscapes. Daytime passive radiative cooling materials shed heat from the ground to the cold universe by taking advantage of the terrestrial thermal radiation that is as large as the renewable solar energy. Newly developed photonic materials permit subambient cooling under direct sunshine, and their applications are expanding rapidly enabled by scalable manufacturing. We review here the recent advancement of daytime subambient radiative cooling materials, which allow energy-efficient cooling and are paving the way toward technologies that harvest the coldness from the universe as a new renewable energy source.

All materials at a temperature above absolute 0 K continuously absorb and emit electromagnetic waves. Thermodynamics thus calls for a process of radiative heat exchange between objects of different temperatures by absorbing and emitting these energy-carrying electromagnetic waves. Looking at radiative energy exchange on the planetary scale, the Earth has a relatively stable surface temperature near 300 K arising from the balanced radiative heat fluxes between the incoming sunlight and the outgoing thermal radiation to the cold universe. The globally absorbed solar radiation has an immense power of 100 PW, and the total outgoing thermal radiation is similarly enormous (*1*). Using the coldness of the universe as a renewable energy source is therefore of great interest and could be as important as solar energy utilization.

Terrestrial passive radiative cooling harvests the coldness of the universe by exploiting the balance of radiative heat flow. The temperature of a sky-facing object can fall below the ambient air temperature by minimizing its heat gain while maximizing the outgoing thermal radiation through the atmospheric window, i.e., 8 to 13 μm. This is the wavelength range at which electromagnetic radiation penetrates the Earth's atmosphere with low losses and a major channel through which the Earth dissipates heat gained from solar to outer space. Unlike almost all other available energy technologies, in which the waste heat is deposited into the surrounding ambient, passive radiative cooling sends excessive heat to the outer space at no additional energy cost (*2*, *3*). Without solar heat, nighttime radiative cooling has been exploited for cooling buildings and for harvesting and purifying water, especially in tropical and subtropical regions (*3*). An understanding of nocturnal radiative cooling is also crucial in applied meteorology and in agriculture to predict frost and to protect fruits or seedlings from freezing (*4*).

[1]Department of Mechanical Engineering, University of Colorado, Boulder, CO 80309, USA. [2]Materials Science and Engineering Program, University of Colorado, Boulder, CO 80309, USA. [3]School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China. [4]Department of Civil and Architectural Engineering, University of Wyoming, Laramie, WY 82071, USA. [5]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA.
*Corresponding author. Email: xiaobo.yin@colorado.edu (X.Y.); ronggui@hust.edu.cn (R.Y.)

"Terrestrial passive radiative cooling harvests the coldness of the universe by exploiting the balance of radiative heat flow."

Daytime radiative cooling, however, poses a fundamental material challenge. To achieve a surface temperature below the ambient temperature under direct sunlight, it is necessary to ensure that the peak solar absorbance does not exceed the energy amount emitted from the surface. Because of the large power density mismatch between terrestrial solar irradiation and infrared thermal radiation from a surface with a temperature close to ambient, the material has to have a very low absorptivity for any wavelengths at which solar energy is not negligible (0.3 to 4 μm) while also having a high absorptivity and emissivity at mid-infrared wavelengths, preferably within the atmospheric window. This spectroscopic requirement is rather stringent because it requires tailoring of optical properties of a material over a broad range of wavelengths that span nearly two orders of magnitude. Natural materials are not readily available for effective daytime subambient cooling.

Experimental demonstrations of daytime subambient cooling with engineered photonic structures have therefore recently attracted great interest (*5*–*7*). Passive radiative cooling at no cost of electricity offers vast opportunities for energy-saving applications in buildings, transportation, cold chain industries, and personal comfort for outdoor activities. Integrating radiative cooling with cold collection and storage can actively improve the efficiency of air-conditioning (*8*, *9*). Although the net cooling energy flux is limited, terrestrial radiative cooling has emerged as a promising solution for mitigating urban heat islands and for potentially fighting against global warming if it can be implemented at a large scale (*3*). The transformative concept of using the universe as a cold sink in an energy system redefines the thermodynamic limits for energy-harvesting processes on Earth (*10*) and has led to emerging effects such as generating light from darkness (*11*).

Fundamentals of radiative cooling

The atmosphere sets the fundamental limit for terrestrial radiative cooling. It partially absorbs and reflects the incident solar radiation. In addition, water vapor and a variety of other greenhouse gases such as carbon dioxide in the air strongly absorb the long-wavelength terrestrial thermal radiation and prevent the heat from being directly released into the cold space. By Kirchhoff's law, such an absorption results in a strong downward radiation from the atmosphere to the terrestrial surface. The net energy flux from terrestrial surfaces is therefore the combination of the absorption of solar radiation (during the day), the outgoing terrestrial thermal radiation, and the absorption of the downward atmospheric thermal radiation (Fig. 1A). The large power density mismatch between the terrestrial solar irradiation and the infrared thermal radiation clearly requires a daytime radiative cooling material to have a very low solar absorptivity (Fig. 1B). For a clear sky, the terrestrial radiation with wavelengths within the atmospheric window is highly transmissive to the entire atmosphere, and the downward atmospheric thermal radiation is insignificant at these wavelengths. Conversely, the terrestrial thermal radiation outside of the atmospheric window can barely escape the atmosphere. The downward thermal radiation of the atmosphere is thus primarily at these wavelengths. Designing a surface that absorbs neither solar radiation nor downward atmospheric radiation is essential to achieving efficient radiative cooling, for example, reaching a deep subambient temperature.

The distinctive spectroscopic feature of the atmospheric window and the possible absorption from the downward atmospheric radiation led to two types of designs for daytime radiative cooling materials: (i) a near blackbody broadband design with a close-to-unity emissivity for all infrared wavelengths and (ii) a wavelength-selective design that has a high emissivity only within the atmospheric window. For both cases, the absorptivity has to be near

zero in the entire solar spectrum. The broadband design emits strongly in all infrared wavelengths. However, because of the reciprocity of radiation and absorption, the large emissivity at the wavelengths outside of the atmospheric window leads to a strong absorption of the downward atmospheric radiation. This adverse effect limits the lowest temperature that is achievable by a broadband emitter. When a broadband emitter is located in a complex environment receiving radiation from surrounding structures such as the ground and buildings, the effectiveness of cooling will be further compromised (*12*). By contrast, the wavelength-selective design has a low emissivity at wavelengths outside of the atmospheric window. The low emissivity reduces the overall power density of outgoing thermal radiation but it also stops the absorption of the downward radiation from the atmosphere and its surroundings, allowing a lower achievable temperature below the ambient temperature. The inset of Fig. 1C shows a nearly linear response of the net radiative cooling power for the two ideal emitters as a function of the temperature difference between the emitter surfaces and their surroundings. Clearly, the broadband cooler has a larger net cooling power when the surface temperature is above or close to the surrounding ambient temperature, which is particularly useful for applications with above-ambient temperature radiant surfaces. However, the wavelength-selective cooler allows achieving a much lower subambient temperature. The nonradiative heat transfer processes of convection and conduction occurring at surfaces also substantially affect the overall radiative cooling performance.

The spectral power density of the downward atmospheric radiation is highly sensitive to the geoclimatic conditions. Putting solar absorbance aside, ground temperature and precipitable water in the air are the two most important factors affecting the net radiative cooling power (Fig. 1D). The transparency of the atmospheric window decreases with the amount of precipitable water, and radiative cooling is therefore commonly considered to be more effective in hot and arid regions. Even under 100% relative humidity in humid areas, experiments have demonstrated >5°C subambient cooling using a wavelength-selective radiative cooling surface (*13*). Other major greenhouse gases such as carbon dioxide and ozone, along with atmospheric particulate matter with a diameter of <2.5 μm (PM2.5) from pollution or other processes, also affect the infrared absorption spectra (Fig. 1D) (*14*). Radiative forcing factors such as increasing greenhouse gas emissions and air pollutants clearly can strongly negate the transparency of the atmospheric window and reduce the effectiveness of terrestrial radiative cooling.

Deep subambient daytime radiative cooling

A large variety of organic and inorganic materials are broadband emissive in the infrared wavelengths and are all capable of pumping heat off terrestrial surfaces at night. For example, a millimeter-thick plate of magnesium oxide or lithium fluoride can achieve 20°C subambient cooling during a dry and clear night (*15*). Increasing solar albedo, i.e., specular or diffusive reflectance of sunlight, is the key to achieving subambient daytime radiative cooling. A wide range of materials with high solar reflectivity have been explored over the past several decades. However, achieving subambient cooling under direct sunlight has been quite a challenge (*5*, *7*). Since the experimental demonstration of daytime subambient cooling (4), a large number of materials have emerged, of which many are broadband emitters (*6*, *16*). For example, by further engineering commercially available coextruded polyester multilayer mirrors and cool roof paints, a multilayer mirror achieved a 2°C subambient cooling without a convection shield (*6*), and the latter could reach subambient temperature for most of the day with a polyethylene convection shield (*16*). While daytime subambient cooling is not yet achievable in natural materials, several examples of effective radiative cooling are known. For instance, the nanostructured hairs from silver ants of the Saharan desert have relatively low solar absorbance (<40%) but also have a high, broadband infrared emissivity (Fig. 2A) (*17*). This allows the silver ant to shed off solar radiation and body heat to thermoregulate its temperature. White silk cocoons also have an average solar absorbance as low as 10% (Fig. 2A) (*18*, *19*). Further processing of these natural materials can lead to synthetic

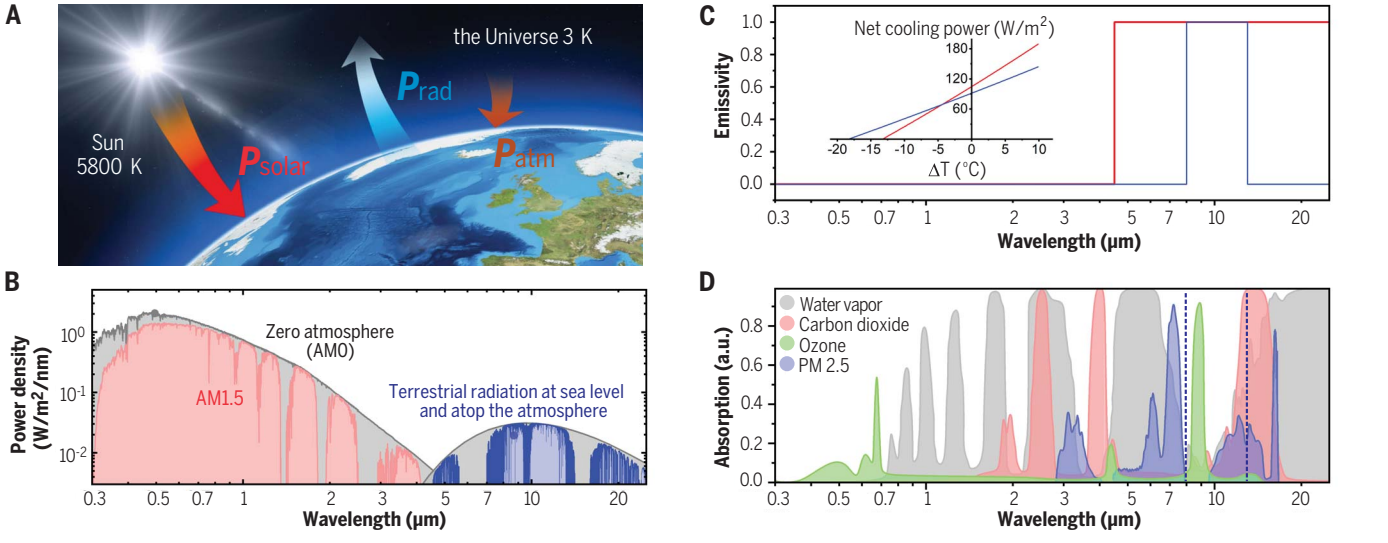


Fig. 1. Fundamentals of terrestrial radiative cooling. (**A**) Schematic of radiative heat flows occurring at terrestrial surfaces. P_{solar}, absorption of solar radiation; P_{rad}, outgoing terrestrial thermal radiation; P_{atm}, absorption of downward atmospheric thermal radiation. (**B**) Solar spectra and terrestrial radiation spectra, respectively, at sea level and atop the atmosphere. Adapted from (*7*). (**C**) Ideal emissivity spectra of the broadband (red) and the wavelength-selective (blue) radiative cooling surface. Inset shows the net cooling power as the function of the temperature difference from the ambient assuming an overall heat transfer coefficient of 3 W/m²•K, a ground ambient temperature of 20°C, and a total precipitable water of 10 mm. Adapted from (*3*) with permission of AIP Publishing. (**D**) Absorption spectra of the major greenhouse gases of water vapor (gray), carbon dioxide (red), and ozone (green), as well as the particulate matters collected from air pollution. The spectrum for particulate matters was adapted from (*14*) with permission from Elsevier.

materials for daytime subambient cooling. For example, delignified natural wood has a wavelength-dependent emissivity (Fig. 2A, blue curve) that allows >4°C subambient cooling during the middle of the day (*20*).

Subambient cooling under direct sunlight poses a photonic design challenge. By tailoring structures at wavelength and subwavelength scales, the rapidly evolving field of nanophotonics provides a means to precisely manipulate materials' electromagnetic properties. Demonstrations involving photonic crystals and metamaterials show how they can control the emissivity of a material or a device (*21*). By stacking two thermally emitting photonic crystals together, Rephaeli *et al.* introduced a wavelength-selective emitter for deep subambient cooling (Fig. 2B, inset) (*22*). The top photonic crystal was designed to enhance the infrared emissivity only within the atmospheric window while being transparent to the rest of the spectrum (Fig. 2B). A broadband chirped design was used for the photonic crystal underneath to maximize the reflectance over a wide range of solar wavelengths. Both functions were subsequently achieved to a certain extent using a single photonic multilayer structure (Fig. 2C, inset) made of alternative layers of silicon dioxide (SiO_2) and hafnium dioxide (HfO_2) (*5*). When back coated with a silver mirror, the multilayer structure reflects nearly 97% of solar radiation while maintaining a relatively strong emissivity within the atmospheric window (Fig. 2C). This design allows for an impressive subambient temperature of 4.9°C (Fig. 2D) when exposed to 850 W/m^2 of direct sunlight. A variety of photonic designs that resonantly enhance the light-matter interactions within the atmospheric window have subsequently been developed. Most incorporate lossy dielectric materials such as heavily doped semiconductors (*2*, *23*) or materials with photon polariton resonances within the atmospheric window (e.g., SiO_2 and silicon carbide) (*3*, *24*). We highlight two photonic nanostructures with excellent wavelength-selective emissivity within the atmospheric window (Fig. 2, E and F). Computational methods that use learning and optimization algorithms have been adopted for the automated design of radiative cooling materials (*24*, *25*). Unfortunately, many of these outstanding photonic structures either remained as theoretical designs or have not been demonstrated at the large scales required for the likely applications. The inherently low energy flux of daytime passive radiative cooling, typically around 100 W/m^2 or less, demands a macroscopic implementation of these microfabricated and/or nanofabricated photonic structures. The large-scale precision lithography and the demanding process of depositing multiple dielectric films that are often involved in these photonic designs clearly hinder their wide deployment.

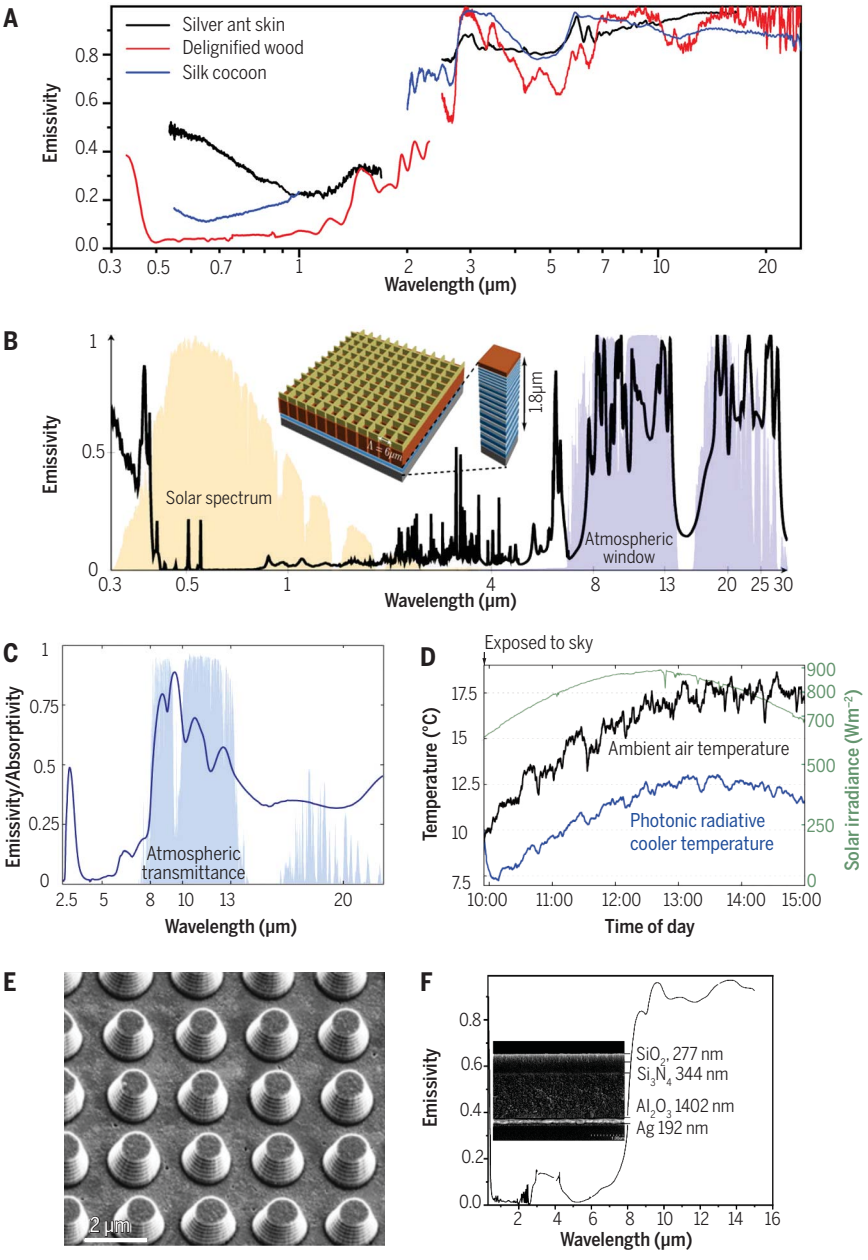


Fig. 2. Photonic materials for radiative cooling. (**A**) Emissivity spectra of natural materials and synthetic materials made from natural resources. Black, skins of silver ants (*17*); blue, silk cocoon (*18*); red, delignified wood (*20*). (**B**) Emissivity spectrum of an optimized photonic structure with deep subambient radiative cooling capacity (inset), which consists of two thermally emitting photonic crystals. Reprinted with permission from (*22*). Copyright 2013 American Chemical Society. (**C**) Infrared emissivity spectrum of a multilayer radiative emitter (inset) showing strong emissivity within the atmospheric window. Reprinted from (*5*) with permission from Springer Nature. (**D**) Midday ambient temperature and surface temperature of the multilayer radiative emitter when solar irradiance is between 800 and 870 W/m^2. Reprinted from (*5*) with permission from Springer Nature. (**E**) Wavelength-selective radiative emitter with multilayer conical pillar arrays of metal-dielectric metamaterial. Reprinted from (*23*) with permission from John Wiley and Sons. (**F**) Multilayer radiative emitter enabled by materials with phonon polariton resonances within the atmospheric window. Reprinted with permission from (*24*). Copyright 2020 American Chemical Society.

Scaling up for application

To make an impact outside of a laboratory-based setting, energy materials need to be scalable. Structural relaxation of an originally strictly defined photonic crystal or metamaterial does not necessarily degrade its performance and, in some cases, can potentially lead to more advantageous properties. Embracing the amorphous nature of random photonic

structures has an obvious advantage in leveraging state-of-the-art manufacturing techniques. For instance, random pyramidally textured silicon surfaces allow high-efficiency light trapping and have been widely adopted in solar cells (*26*). Taking advantage of the well-established roll-to-roll polymer extrusion process, Zhai *et al.* demonstrated a scalable-manufactured metamaterial thin film for deep subambient cooling by encapsulating randomly positioned SiO_2 microspheres in a visibly transparent matrix made of a polymethylpentene polymer (Fig. 3A) (*7*). With the right size of dielectric microspheres, the collective excitation of the surface phonon polariton resonances of the embedded microspheres resulted in a close-to-unity emissivity for the atmospheric window. When coated with a silver mirror underneath (Fig. 3B), the flexible metamaterial thin film achieved a >8°C subambient cooling at noon with >800 W/m^2 of solar irradiance (Fig. 3C). The temperature difference was as large as 15°C when there was no direct sunlight impinging onto the metamaterial surface (Fig. 3D). Scalable manufacturing of the deep subambient radiative cooling metamaterials rapidly enables the implementation of a 13.5-m^2 kilowatt-scale system for cold collection and storage (*9*). The introduction of polymers substantially improved the manufacturability and applicability of radiative cooling materials. Polymer-based radiative cooling materials have now received markedly increased interest. In fact, polymers are understood to be strongly emissive and absorptive in the infrared because of vibrational excitations of functional groups (*28*). For example, the polar carbon-halogen bonds have their fingerprint (bending) vibration frequencies well overlapped with the atmospheric window, making many durable halogen compounds such as polyvinylidene difluoride (PVDF) and polytetrafluoroethene (PTFE) or their blends great candidates for outdoor use of passive radiative cooling (*28*–*31*). We show two polymer-based radiative cooling materials (Fig. 3, E and F) and both have the potential to be scaled up in manufacturing. Instead of using a reflective metallic mirror or a multilayer photonic structure to reject the full spectrum solar radiation, highly hierarchical and random porous networks (Fig. 3E) (*30*) and porous fibrous structures (Fig. 3F) (*31*) were introduced to reject sunlight through multiple scattering. With further development of phase-inversion chemistry using a more eco-friendly solvent, the approach can provide the versatility that is needed to apply radiative cooling materials on existing surfaces and structures.

Polymeric and random photonic materials (Fig. 3, A to F) achieve the needed wavelength selectivity by randomly incorporating the infrared resonators with phonon polariton resonances and/or blending polymers with selected

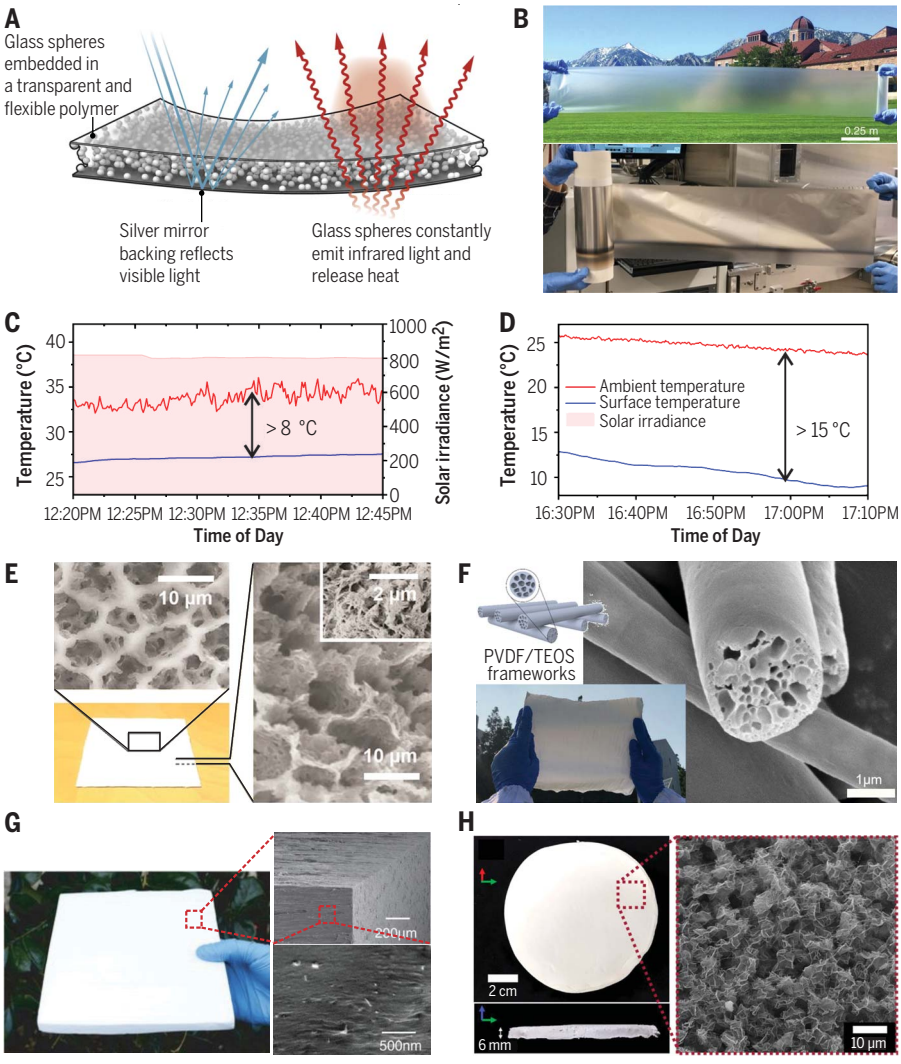


Fig. 3. Scale-up of the subambient radiative cooling materials. (**A**) Schematic of a radiative cooling polymer-inorganic metamaterial consisting of randomly distributed infrared microsphere resonators (*27*). (**B**) Photos of the scalable-manufactured metamaterial thin films before (top) and after (bottom) silver coating. Reproduced from (*3*) with permission of AIP Publishing and from (*7*). The metamaterial thin film shows >8°C subambient cooling at noon with >800 W/m^2 solar irradiation (**C**) and >15°C subambient cooling when there is no direct sunlight received (**D**). Two subambient cooling materials made of porous polymer networks [(**E**) (30)] and fibrous structures [(**F**) Reprinted from (31) with permission from John Wiley and Sons]. (**G**) Delignified and then condensed synthetic wood showing both high toughness and subambient cooling capacity (*20*). (**H**) A polyethylene aerogel serves as a solar reflector while providing thermal isolation (*33*).

vibrational frequencies that are well within the atmospheric window. These have shown compelling advantages in scaling up the fabrication process, and several prototype radiative cooling systems have already been deployed for daytime use (*8*, *9*). The stringent requirement of high solar reflectance has been achieved by backing the material with a reflecting metallic mirror (Fig. 3, A and B) or by multiple scattering from low-loss porous networks (Fig. 3, E and F). Facile synthesis or inclusion of randomly distributed light-scattering microparticles or nanoparticles such as titanium dioxide and aluminum oxide (Al_2O_3) also provides a versatile approach to render an originally transparent material highly reflective in the solar spectrum; however, most of these materials are broadband emitters because of the inclusion of infrared emissive light scatters at high concentrations (*6*, *16*, *32*). Taking advantage of random photonics, multifunctional materials with daytime radiative cooling capability are of great interest. A densified radiative cooling structural material can be made from a delignified natural wood (*20*) (Fig. 3G). The processed wood has a toughness >10 times stronger than that of common natural wood while being capable of subambient cooling throughout the day. By contrast, we present a lightweight material made of a polyethylene

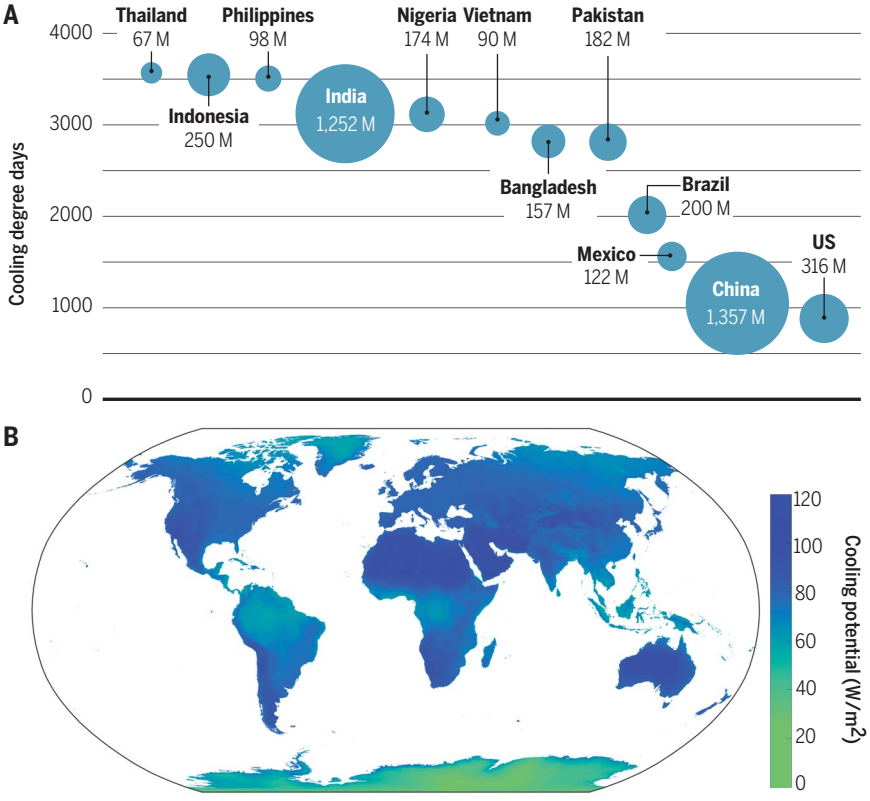


Fig. 4. Air-conditioning potential in selected countries and worldwide radiative cooling potential. (**A**) Twelve counties with large air-conditioning potentials (*35*). The populations of these countries are also labeled. (**B**) Global radiative cooling potential.

aerogel (*33*) (Fig. 3H). The high porosity makes the aerogel highly solar reflective with ultralow thermal conductivity. Moreover, the aerogel is spectroscopically transparent in the infrared. When applying the aerogel on a broadband emitter, the combined structure demonstrated a strong subambient cooling effect because the aerogel not only allows the infrared thermal radiation to transmit through but also provides an excellent thermal isolation for reducing heat loss from the emitter surface. The same strategy of being spectrally reflective in solar wavelengths while being transparent in infrared has also been recently exploited to produce cooling textiles for personal thermal management (*34*).

Enabling applications

Heat stress caused by high temperature is one of the most severe climate threats to human health and to the well-being of our society. The use of air-conditioning, along with the associated energy consumptions and environmental costs, is projected to increase substantially in the following decades (*35*) (Fig. 4A). The electricity needed for global air-conditioning is expected to triple over the next two decades, a trajectory that is not sustainable. Direct and passive integrations of radiative cooling materials with building envelopes such as roofs, walls, and windows can result in substantial energy savings, especially for hot and/or arid climates. Active circulation of radiatively cooled air can better regulate the indoor temperature for energy saving in regions with hot summers but have a reduced penalty of heating in cold winters by simply shutting off the circulation (*36*). The wavelength-selective radiative cooling materials that allow deep subambient cooling are of particular advantage for the generation of cold water that can be used to reduce a cooling system's condenser temperature for improved efficiency; a 1°C reduction of condenser temperature may lead to 3 to 5% electricity saving for a chiller (*8*, *9*). Considering the tremendous freshwater withdrawal for power plant cooling, the coldness generated by radiative cooling could reduce water use while increasing power plant efficiency if implemented at a scale that reduces the steam condensation temperature. Similarly, the potential of passive radiative cooling has been investigated in photovoltaics to reduce operation temperature and to increase efficiency; for example, this has been applied in electric automobiles to reduce air-conditioning use and prolong mileage and in consumer electronics for thermal management (*2*, *3*). Untapped coldness can also be used to build a low-temperature environment for thermoelectric generation (*11*) and other thermal cycles (*37*).

To identify the best climates for potential large-scale deployment of terrestrial radiative cooling, in Fig. 4B we show a world map of annual radiative cooling potential using an ideal wavelength-selective radiative cooling material with a 5% solar absorbance. The assessment is based on the global meteorological datasets collected between 2015 and 2019, including rainy and cloudy weather conditions (*38*). The radiative cooling potentials are clearly geographically dependent because of regional climates. Regions with a dry and hot climate tend to have higher radiative cooling power, up to 120 W/m^2. By comparison, cold, highly humid, or cloudy regions tend to have much lower radiative cooling potentials. Wide adoption of radiative cooling materials in regions with high radiative cooling potentials can radically change building architectures because the technology not only improves building heat dissipation in both day and night but also reduces the need for roof insulation or solar shading structures often used in these areas.

Finally, increasing the terrestrial albedo of buildings alone could reduce the heat island effect by 33% for urban landscapes (*39*). Applying passive radiative cooling materials at an even larger scale over pavements and building roofs across a densely populated city offers a transformative geoengineering tool for mitigating urban heat stress, potentially addressing the crisis of global warming (*40*).

Conclusion and Outlook

Radiative cooling materials harvesting the coldness from the universe can be a game changer for reducing cooling demands, increasing power production, and even mitigating the urban heat island effect. This rapidly evolving field has the potential to change the present global energy landscape. For wider adoptions of daytime passive cooling materials and technologies, however, substantial challenges remain that should motivate research efforts. There is a strong need for a reliable, cost-effective, self-adapting, and colorful radiative cooling material. Some progress has been made in this direction. For example, to ensure that the cooling benefits in summer far exceed the penalties in winter, especially in high-latitude areas, switchable and self-adaptive radiative cooling materials were explored with electrochromic and thermochromic materials (*41*, *42*). Radiative cooling materials with colors have been recently explored by minimizing the adverse solar absorbance from colorants (*43*, *44*). By using photocatalytic titanium dioxide as a top coating to accelerate the degradation of organic pollutants, self-cleaning radiative cooling materials can potentially suppress soiling and dust accumulations (*45*). Radiative

cooling harvests the coldness from the universe but cools atop a surface. From a system perspective, maximizing the net radiative cooling capacity also needs to be coupled with enhancing the heat exchange for high-efficiency cold collection while eliminating, if possible, the convective and conductive losses from the radiative surfaces. Moreover, recognizing that the precipitable water and cloud coverage has a profound impact on the effect of radiative cooling (*9*, *13*) is important, and a detailed understanding of such an effect is required for the evaluation of radiative cooling technology. Performance evaluation of radiative cooling materials has been sporadic and complicated by geoclimatic testing conditions. To harvest the coldness from the dark universe as a new renewable energy resource, standardized tests across global and large-scale deployment of pilot systems are still needed.

REFERENCES AND NOTES

1. G. L. Stephens, T. L'Ecuyer, *Atmos. Res.* **166**, 195–203 (2015).
2. M. M. Hossain, M. Gu, *Adv. Sci.* **3**, 1500360 (2016).
3. D. L. Zhao *et al*., *Appl. Phys. Rev.* **6**, 021306 (2019).
4. U. Chung *et al*., *Agric. For. Meteorol.* **137**, 15–24 (2006).
5. A. P. Raman, M. A. Anoma, L. Zhu, E. Rephaeli, S. Fan, *Nature* **515**, 540–544 (2014).
6. A. R. Gentle, G. B. Smith, *Adv. Sci.* **2**, 1500119 (2015).
7. Y. Zhai *et al*., *Science* **355**, 1062–1066 (2017).
8. E. A. Goldstein, A. P. Raman, S. Fan, *Nat. Energy* **2**, 17143 (2017).
9. D. L. Zhao *et al*., *Joule* **3**, 111–123 (2019).
10. W. Li, S. Buddhiraju, S. Fan, *Light Sci. Appl.* **9**, 68 (2020).
11. A. P. Raman, W. Li, S. Fan, *Joule* **3**, 2679–2686 (2019).
12. R. G. Yang, X. B. Yin, *Nat. Sustain.* **2**, 663–664 (2019).
13. M. Dong, N. Chen, X. Zhao, S. Fan, Z. Chen, *Opt. Express* **27**, 31587–31598 (2019).
14. X. Yu *et al*., *J. Environ. Sci. (China)* **71**, 76–88 (2018).
15. P. Berdahl, *Appl. Opt.* **23**, 370 (1984).
16. S. Atiganyanun *et al*., *ACS Photonics* **5**, 1181–1187 (2018).
17. N. N. Shi *et al*., *Science* **349**, 298–301 (2015).
18. S. H. Choi *et al*., *Nat. Commun.* **9**, 452 (2018).
19. N. N. Shi *et al*., *Light Sci. Appl.* **7**, 37 (2018).
20. T. Li *et al*., *Science* **364**, 760–763 (2019).
21. S. Fan, *Joule* **1**, 264–273 (2017).
22. E. Rephaeli, A. Raman, S. Fan, *Nano Lett.* **13**, 1457–1461 (2013).
23. M. M. Hossain, B. H. Jia, M. Gu, *Adv. Opt. Mater.* **3**, 1047–1051 (2015).
24. D. Chae *et al*., *ACS Appl. Mater. Interfaces* **12**, 8073–8081 (2020).
25. K. Kitai *et al*., *Phys. Rev. Res.* **2**, 013319 (2020).
26. P. Campbell, M. A. Green, *J. Appl. Phys.* **62**, 243–249 (1987).
27. X. Zhang, *Science* **355**, 1023–1024 (2017).
28. A. Aili *et al*., *Materials Today Physics* **10**, 100127 (2019).
29. P. Yang, C. Chen, Z. M. Zhang, *Sol. Energy* **169**, 316–324 (2018).
30. J. Mandal *et al*., *Science* **362**, 315–319 (2018).
31. X. Wang *et al*., *Adv. Funct. Mater.* **30**, 1907562 (2020).
32. J. Mandal, Y. Yang, N. F. Yu, A. Raman, *Joule* **4**, 1350–1356 (2020).
33. A. Leroy *et al*., *Sci. Adv.* **5**, eaat9480 (2019).
34. P. C. Hsu *et al*., *Science* **353**, 1019–1023 (2016).
35. L. W. Davis, P. J. Gertler, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 5962–5967 (2015).
36. D. L. Zhao, A. Aili, X. B. Yin, G. Tan, R. G. Yang, *Energy Build.* **203**, 109453 (2019).
37. Z. Chen, L. Zhu, W. Li, S. Fan, *Joule* **3**, 101–110 (2019).
38. S. E. Fick, R. J. Hijmans, *Int. J. Climatol.* **37**, 4302–4315 (2017).
39. K. W. Oleson, G. B. Bonan, J. Feddema, *Geophys. Res. Lett.* **37**, L03701 (2010).
40. J. N. Munday, *Joule* **3**, 2057–2060 (2019).
41. J. Mandal *et al*., *Adv. Funct. Mater.* **28**, 1802180 (2018).
42. M. Ono, K. Chen, W. Li, S. Fan, *Opt. Express* **26**, A777–A787 (2018).
43. W. Li, Y. Shi, Z. Chen, S. Fan, *Nat. Commun.* **9**, 4240 (2018).
44. Y. Chen *et al*., *Sci. Adv.* **6**, eaaz5413 (2020).
45. G. Y. Wei, D. Y. Yang, T. Zhang, X. J. Yue, F. X. Qiu, *Sol. Energy Mater. Sol. Cells* **202**, 110125 (2019).

ACKNOWLEDGMENTS

X.Y. gratefully acknowledges support from a Moore Inventors fellowship (grant no. 6884). S.F. acknowledges support from the U.S. Department of Energy (grant no. DE-FG-02-07ER46426).

10.1126/science.abb0971

REVIEW

New refrigerants and system configurations for vapor-compression refrigeration

Mark O. McLinden[1]*, Christopher J. Seeton[2], Andy Pearson[3]

The high global warming potential of current refrigerants in cooling equipment based on the vapor-compression cycle has triggered a major effort to find and implement more environmentally benign alternatives. Here, we review the basics of the vapor-compression cycle together with the safety, environmental, and thermodynamic constraints that have led to the current and next generation of refrigerants. The development of new fluids has focused on fluorinated olefins, known as hydrofluoroolefins (HFOs), and blends that contain HFOs. Many of these are slightly flammable, presenting trade-offs between safety and environmental considerations. Engineers also have options with a resurgence of the "natural refrigerants" (ammonia, carbon dioxide, propane, and isobutane). Innovative system designs that reduce the required quantity of refrigerant may allow a wider choice of refrigerants.

Refrigeration and cooling are pervasive in modern society, from the refrigerated cold chain that provides us fresh meats and vegetables to the air conditioning of homes and offices and the cooling of data servers that enables the internet and global communications. Refrigeration is viewed by some as one of the 20 most transformative engineering achievements of the 20th century (*1*). The demand for cooling will continue to increase for many years because of improving economic conditions globally along with changing regional climate.

Heat normally flows from hot to cold, but refrigeration enables a reversal of that flow to either make something colder (such as refrigeration, air conditioning, or process cooling) or upgrade heat from a cold ambient source (heat pumps). A wide variety of methods and systems exist to provide refrigeration, and the most common type is based on the vapor-compression cycle. The key to these systems is the working fluid, the refrigerant, that transports heat around the cycle through phase changes (boiling and condensation).

Although refrigeration provides tremendous societal benefits, it comes with environmental costs. Refrigeration accounts for 7.8% of total global greenhouse gas emissions, with 63% of that from "indirect" emissions (*2*) stemming from the energy that drives the equipment. This value reflects the environmental impacts of the largely fossil fuel–based energy sector. Focusing on the refrigerant, the switch to fluids with substantially lower direct global warming impact could avert nearly 3% of global greenhouse emissions, with larger impacts in the future as the use of refrigeration and air-conditioning grows.

At the turn of the 20th century, early refrigerants were flammable or toxic, so they were mainly used in industrial systems. Ammonia, widely used in ice-making plants, is toxic and flammable, although difficult to ignite. Smaller systems used toxic sulfur dioxide or flammable and toxic methyl chloride. The demand for safer refrigerants—suitable for use in homes and public spaces—prompted the development of a class of halogenated compounds known as chlorofluorocarbons (CFCs) and hydrochlorofluorocarbons (HCFCs), also called "freons." These compounds are nonflammable and have low toxicity, making them seemingly perfect refrigerants. However, they are long-lived in the atmosphere and transport chlorine to the stratosphere. They also have a high direct global warming impact. The environmental consequences of the CFCs and HCFCs began to be recognized in the 1970s, and chlorine from them was definitively linked to the depletion of stratospheric ozone in 1986 (*3*). In 1987, an international treaty, the Montreal Protocol, first phased down and, with later revisions, ultimately phased out the CFC and HCFC refrigerants.

A related class of chemicals, the hydrofluorocarbons (HFCs), were brought to market in the 1990s as replacements. The HFCs contain no chlorine and thus have no ozone depletion potential (ODP), but they retain the heat-trapping characteristics of the CFCs and HCFCs as quantified by the global warming potential (GWP), which is a measure of the climate impact of a gas relative to that of carbon dioxide (CO_2). The value is calculated over some time horizon, most commonly 100 years, which is denoted GWP_{100}. Values for GWP_{100} vary from less than 1 for some of the new refrigerants to 23,500 (*4*) for SF_6, an insulating gas used in electrical equipment. The GWP_{100} for the most common CFCs and HCFCs range from 1800 to 14,000.

By the early 2000s, the potential climate impacts of the HFCs were beginning to be considered. The HFCs generally have lower GWP than the CFCs and HCFCs that they replaced, and their introduction enabled a relatively rapid transition away from the CFCs, thus averting some climate impacts (*5*). Although current HFC emissions are estimated to contribute less than 0.02 K to warming at the surface, a business-as-usual scenario with increasing emissions would lead to a more substantial climate impact; one estimate by Velders *et al.* is from 0.3 to 0.5 K by the end of the century (*6*, *7*).

A further amendment to the Montreal Protocol, known as the Kigali Amendment, was negotiated in 2016. It phases down the high-GWP HFCs, calling for an eventual reduction of 85% in GWP-weighted production by 2036 in developed countries, with later dates in other parts of the world. The details of this phase down are left to each country. These are promulgated by the Environmental Protection Agency (EPA) Significant New Alternatives Program (SNAP) regulations in the United States and the "F-Gas" regulations of the European Union (EU).

Once again, the refrigeration industry is searching for replacements for the fluids vital to the operation of its equipment. We survey current developments to find and implement replacement refrigerants, discuss the resurgence of long-known natural refrigerants, and review system innovations that would allow the use of alternative refrigerants or otherwise reduce environmental impacts.

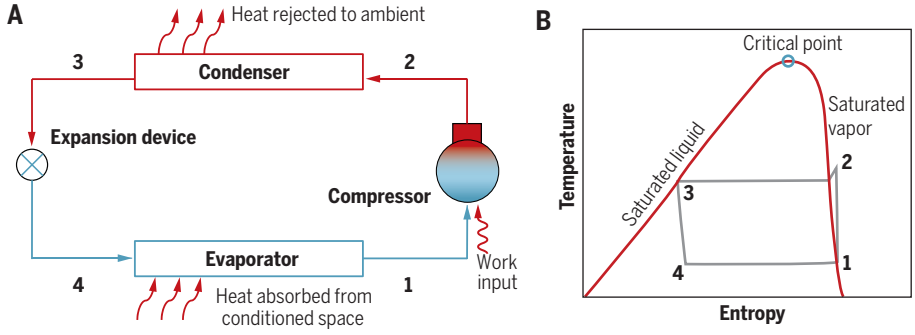


Fig. 1. Vapor-compression cycle. Vapor-compression refrigeration systems operate by boiling a working fluid (the refrigerant) in the evaporator (4 → 1) at a relatively low pressure, extracting heat from a low-temperature source (such as the refrigerated space). The resulting vapor is compressed to a higher pressure (1 → 2). It is then condensed back to a liquid (2 → 3), with the heat of condensation rejected to a relatively high-temperature heat sink (ambient air for most refrigeration and air-conditioning systems). The refrigerant is then reduced back to the starting pressure to complete the cycle (3 → 4). (Right) The cycle on a thermodynamic diagram. The condensation and evaporation processes are (ideally) at constant pressure and, thus, constant temperature (for a pure fluid); the ideal compression process is at constant entropy.

Basics of vapor-compression refrigeration

Vapor-compression refrigeration systems exploit two fundamental properties of all fluids: (i) The boiling temperature varies with pressure, and (ii) a change in phase (liquid boiling to a vapor and condensing back to a liquid) is accompanied by the absorption or release of heat. A common example is that the boiling temperature of water is raised above 100°C as pressure increases in a pressure cooker. Likewise, the boiling temperature of a refrigerant varies with pressure, and by manipulating the pressure of a refrigerant at different points in a refrigeration cycle, it can boil at a low pressure and low temperature while absorbing heat. By compressing the resulting refrigerant gas to a higher pressure, it condenses back to a liquid and releases the heat at a higher temperature (Fig. 1). Depending on the desired effect, the basic cycle can be used either to heat an object (heat pump heating) or cool an object (air conditioning and refrigeration).

The refrigerant is not consumed in this process, but cycled indefinitely. Only when refrigerant escapes into the atmosphere through leaks or poor technique in service and disposal does the ODP or GWP of a refrigerant actually affect the environment. Thus, the direct GWP of a refrigerant resulting from emissions of the refrigerant to the atmosphere is distinct from the indirect GWP effects resulting from the energy required to drive the cycle, such as CO_2 emissions from a power plant burning a fossil fuel. The ratio of indirect to direct effects varies with the refrigerant, application, and energy supply of a region. With proper service practices that minimize leaks, the indirect effects are dominant in most cases.

Vapor-compression systems are used in a wide range of devices, from small refrigerators to massive central systems that provide many megawatts of cooling capacity to serve an entire university campus, office park, or industrial complex. The compressors fall into three broad categories: (i) reciprocating compressors (analogous to a typical automobile engine) and other types of linear compressors; (ii) rotary compressors, including screw, scroll, rolling piston, spool, and other single- or multivane types, in which the pressure rise results from progressive reduction in the volume of cavities within the machine on a continuous basis; and (iii) kinematic compressors, including radial and axial turbines (analogous to jet engines), which impart kinetic energy to a flow of gas; the kinetic energy is then converted to a pressure rise. The basics of compressor technology are mature in general, but innovative system designs can allow the different types of compressors (and the corresponding types of refrigerants) to be used in a wide variety of new applications.

Vapor-compression systems dominate refrigeration in both numbers and refrigeration capacity. Reasons for this dominance include the maturity of the technology, ease of manufacturing, and scalability from small to very large systems. The use of an evaporating fluid to provide the cooling effect is a convenient method of moving heat in a compact and effective way compared with most alternative technologies.

Constraints—How did we get here?

The desired and required criteria of a refrigerant have been laid out over the years (Box 1) (*8*–*10*). Many of these criteria are not strictly required to realize a functioning refrigeration system but are mandated by codes, standards, and regulations. Standards define best practices and are usually voluntary. Codes define required practices and are enforced by various government agencies. Codes often incorporate the practices defined in standards. The most widely recognized refrigeration safety standard in the United States is American National Standards Institute (ANSI)/American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Standard 15 (*11*), which is codified in the electrical and mechanical codes adopted by local or state governments. The equivalent international standard is International Organization for Standardization (ISO) 5149 (*12*). Environmental criteria are imposed by national or local regulations, often in response to international agreements, such as the Montreal Protocol.

Box 1. Requirements of a refrigerant.

Chemical

Stable for the life of the system

Health and safety

No/minimal flammability
Low toxicity

Environmental

Zero/de minimis ODP
Low GWP
Minimal secondary environmental impacts (for example, water and local air pollution)
Maximum energy efficiency

Thermodynamic

Properties matched to the application and system

Product sustainability

Long operational life
Maximize recyclable content
Minimize material use

Practical

Materials compatibility
Reasonable cost

The refrigerants themselves are classified in the parallel ANSI/ASHRAE Standard 34 (*13*) and ISO Standard 817 (*14*). The safety classification consists of a letter indicating toxicity, with "A" indicating "lower toxicity" and "B" indicating "higher toxicity." The flammability classification is "1," "2L," "2," or "3," ranging from 1 for refrigerants exhibiting "no flame propagation" in a standardized test (such as the CFCs and many of the HFCs) to 3 for "higher flammability" (such as hydrocarbons). Of special relevance to some of the newer refrigerants is the "A2L" classification, which indicates lower toxicity and lower flammability. Standard 34 and ISO 817 also define a standard nomenclature for refrigerants (Fig. 2).

Over the years, the desired criteria have become increasingly stringent. Before the 1930s, refrigeration systems were largely in industrial settings, such as ice plants and cold stores, and ammonia emerged as the preferred refrigerant. The availability of the nonflammable, low-toxicity CFCs and HCFCs opened up new markets in domestic and commercial refrigeration and air-conditioning, which quickly grew to surpass the industrial refrigeration market. Calm reviews the history of refrigerants and characterizes this period as "safety and durability" (*15*). The HFCs are part of the "third generation" of refrigerants, characterized by Calm as "ozone protection" (*15*).

Faced with the mandate to phase out the CFCs and HCFCs and the newer mandate to phase down the HFCs, four primary approaches exist to find replacements: (i) a "design-compatible" replacement, which involves finding a new fluid with thermodynamic properties similar to those of existing refrigerants while avoiding undesirable characteristics, thus enabling use of existing technologies with minimal retooling and disruption; (ii) selecting a fluid based on environmental (or other) properties and redesigning equipment as necessary to deal with operational changes (such as a higher operating pressure), possible hazards, and materials compatibility issues; (iii) implementing new thermodynamic cycles, including separating the heat load from the latent load (associated with dehumidification); and (iv) implementing a completely different technology, such as thermoelectric devices [He and Tritt (*16*)] or thermocaloric cooling systems (*17*). Brown and Domanski provide a summary that outlines the range of alternative cooling technologies (*18*). We discuss the first two of these primary approaches in this Review.

Thermodynamic requirements of a refrigerant

Refrigeration cycles are governed by thermodynamics. The thermodynamic properties of a

[1]Applied Chemicals and Materials Division, National Institute of Standards and Technology, Boulder, CO 80305, USA. [2]Shrieve Chemical Products, The Woodlands, TX 77380, USA. [3]Star Refrigeration, Glasgow, Scotland, UK.
*Corresponding author. Email: mark.mclinden@nist.gov

fluid include its vapor pressure, density, and heat capacity, and these determine the operating pressures of a refrigeration system along with having a substantial impact on the energy efficiency. The nature of the vapor-compression cycle requires a volatile fluid, and these will be small molecules (typically based on one to four carbons), with boiling points generally from −50° to +30°C.

The search for the ideal refrigerant has been a recurrent theme throughout the history of mechanical refrigeration. There exists, however, no single ideal refrigerant. The desired properties of the refrigerant depend on the type of equipment and the operating conditions (primarily heat sink and source temperatures). McLinden *et al.* carried out a systematic search for suitable refrigerants (*19*–*21*). The key thermodynamic parameters were the liquid-vapor critical temperature and pressure and the heat capacity of the vapor. The results indicated limited options for new refrigerants, and most of the low-GWP options were at least slightly flammable. The search identified many of the newer fluids already under consideration but only added a few candidates, most of which had unknown hazards.

New refrigerants

Most refrigerating systems in the world use fluorocarbon refrigerants. Faced with required changes, a transition to fluids with similar thermophysical properties enabled redesign and retooling costs to be minimized and hence facilitated a rapid response to the restrictions on CFCs and HCFCs. HFCs were established as the preferred replacement within a decade, and the Antarctic ozone layer is recovering (*22*). However, even early in the development of these replacements, their high GWP was recognized as an issue. For example, in 1993, the automotive industry underwent a 2-year complete transition from CFC-12 (dichlorodifluoromethane) to HFC-134a (1,1,1,2-tetrafluoroethane). This was by far the fastest transition of any refrigeration application. But HFC-134a's GWP_{100} of 1300 (*4*), although lower than that of R-12, was still above limits set by the 1992 United Nations Framework Convention on Climate Change. In 1995, work began to move away from HFC-134a to CO_2 or another low-global-warming refrigerant, and in 2006, the European Union passed Directive 2006/40/EC (the MAC Directive), which mandated that by 2011, new vehicle models must switch to a refrigerant with $GWP_{100} < 150$.

Hydrofluoroolefins (HFOs) were identified as the most promising replacements. "Olefin" refers to a class of organic chemicals with a double bond between carbon atoms. The double bond is highly reactive to naturally occurring hydroxyl radicals in the atmosphere, resulting in atmospheric lifetimes of days to weeks for the HFOs, compared with years to decades for the HFCs. Thus, the GWP_{100} for most of the HFOs are under 1 (*4*). The HFOs are a subset of the HFCs, but their environmental characteristics are so different that they are almost universally referred to by the HFO label. The HFOs have been known to the chemical industry since the 1930s and were investigated early on as possible refrigerants. On the basis of numerous patents, the majority of the research was centered around their role as precursors to fluoropolymers.

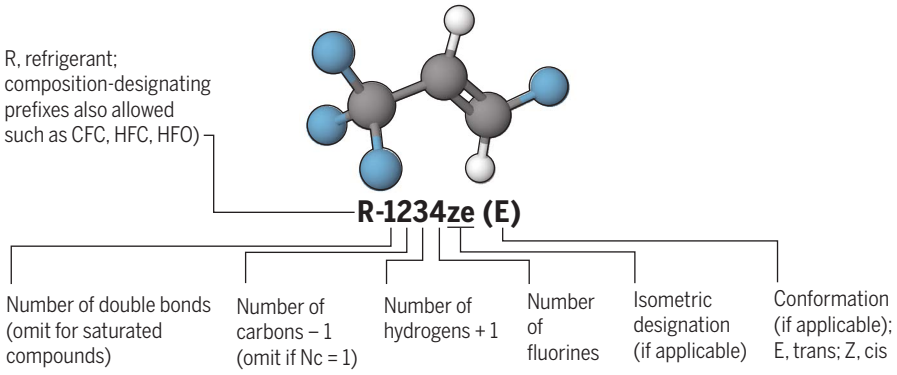

R, refrigerant; composition-designating prefixes also allowed such as CFC, HFC, HFO)
R-1234ze (E)
Number of double bonds (omit for saturated compounds)
Number of carbons − 1 (omit if Nc = 1)
Number of hydrogens + 1
Number of fluorines
Isometric designation (if applicable)
Conformation (if applicable); E, trans; Z, cis

Fig. 2. Alphanumeric soup—Refrigerant nomenclature. ASHRAE Standard 34 and ISO Standard 817 define a nomenclature ("R-numbers") for refrigerants. Here, the HFO *trans*-1,3,3,3-tetrafluoropropene is taken as an example. Additional rules apply for refrigerant blends and inorganic molecules (for example, CO_2 is R-744).

Among the HFOs, those based on the three-carbon propene (propylene) have boiling points in a suitable range for refrigerants and have been the focus of development efforts. As the hydrogens on propene are substituted with fluorine, the flammability decreases. Propene-based HFOs with three or fewer fluorines are flammable, and those with five or six are nonflammable. A pentafluoropropene isomer [HFO-1225ye(Z), *cis*-1,2,3,3,3-pentafluoropropene] was investigated in the Cooperative Research Program of SAE International in 2006 to 2007 (*23*), but there were toxicity concerns. The fully fluorinated HFO-1216 (hexafluoropropene) is also toxic.

Attention then turned to isomers of tetrafluoropropene, and specifically HFO-1234yf (2,3,3,3-tetrafluoropropene). This fluid exhibits some flame propagation in a standardized test and so has a safety classification of A2L under ANSI/ASHRAE Standard 34 (*13*). In practice, HFO-1234yf is difficult to ignite, and its flame is unstable and easily extinguished. HFO-1234yf is now used in a majority of new automotive air-conditioning systems and is expected to be nearly universally used in the coming years in the EU, United States, and Japan. This comes as HFC-134a use in automotive systems is mandated to be phased out in these regions by 2017, 2021, or 2023, respectively (*24*). Most of Asia, including China, continues to use HFC-134a. The progression of refrigerants over time in automotive air conditioning is illustrated in Fig. 3. Another isomer of tetrafluoropropene [HFO-1234ze(E), *trans*-1,3,3,3-tetrafluoropropene] has been commercialized primarily as a foam-blowing agent, although it is also used as a refrigerant in chillers (*24*).

The four-carbon HFO-1336mzz(Z) (*cis*-1,1,1,4,4,4-hexafluorobutene) has been commercialized as a foam-blowing agent and chiller refrigerant [often as a component of a blend with HCO-1130(E) (*trans*-1,2-dichloroethene), known as R-514A (*13*)]. HFO-1336mzz(Z) is nonflammable and has low toxicity but operates at a pressure too low for many types of refrigeration systems. The related isomer HFO-1336mzz(E) (*trans*-1,1,1,4,4,4-hexafluorobutene) is also being studied (*25*). Other HFOs that are being developed include HFO-1132a (1,1-difluoroethene), which is used as a feedstock in fluoropolymers and has a very low boiling point of −83°C (*13*). It would be used in low-temperature applications or to increase the pressure of blends.

HFO-1123 (1,1,2-trifluoroethene) is an unstable molecule that is subject to a disproportionation reaction at high temperatures and pressures. Nevertheless, one company is pursuing it as a replacement for R-410A, arguing that the hazard could be mitigated by blending it with other refrigerants, such as HFC-32 (difluoromethane) (*26*, *27*).

Other olefins contain both fluorine and chlorine and are termed hydrochlorofluoroolefins (HCFOs). These have a small but nonzero ODP. However, they also have extremely short atmospheric lifetimes and very small GWP. HCFO-1233zd(E) (*trans*-1-chloro-3,3,3-trifluoropropene), with ODP < 0.0004 (*28*), has been commercialized as a refrigerant in chillers (and also as a foam-blowing agent) (*24*). HCFO-1224yd(Z) (*cis*-1-chloro-2,3,3,3-tetrafluoropropene) are being classified under ANSI/ASHRAE Standard 34, and its thermodynamic properties are being studied (*25*).

The working fluid in most refrigeration systems is actually a mixture of the refrigerant and a lubricant, which provides lubrication to the moving parts of the compressor. The

oil migrates throughout the system and thus must be miscible with the refrigerant. The HFO refrigerants use similar lubricants as those of HFC refrigerants, namely polyolesters (POEs), polyvinylethers (PVEs), polyethers (PEs), and polyalkylene glycols (PAGs). Each have different application advantages and disadvantages depending on usage, temperature, and equipment. For the HCFOs, the added chlorine in the molecule allows for use of mineral oils (MOs), poly-α-olefins (PAOs), and alkylbenzene (AB) lubricants, as well as the POEs, depending on the application.

None of the HFOs, however, are a direct replacement for R-410A, which is a blend of HFC-32 and HFC-125 (pentafluoroethane) and is used in the majority of small air-conditioning systems. In turn, R-410A was the replacement for HCFC-22 (chlorodifluoromethane) in many applications. ASHRAE Standard 34 list numerous blends that have been proposed as a means to tailor thermodynamic properties to match an existing fluid to enable a direct substitution in existing equipment or to obtain a nonflammable blend. These often contain some combination of HFO-1234yf, HFO-1234ze(E), HFC-134a, and HFC-125; many of these are summarized by Kujak and Schultz (*10*, *29*). To suppress flammability, relatively high concentrations of HFC- 134a and/or HFC-125 are required, and these blends have GWP_{100} values that range from 540 to more than 2000. These numbers are too high to meet the eventual 85% phase-down target of the Kigali Amendment. Many of these blends are intended for retrofit applications, and although they can be considered only as interim solutions, they do enable a lower GWP in existing equipment.

A second approach is to accept a 2L flammability classification in exchange for a much reduced GWP (GWP_{100} of ~100 to 500). However, these fluids require additional safety measures for implementation in the marketplace. HFC-32 makes up a substantial fraction of many of these blends. This approach has been used, for example, to mimic the properties of R-410A.

The carbon-carbon double bond in the HFOs results in a much-reduced atmospheric lifetime, but this increased reactivity can have other consequences. The stability of the HFOs within the sealed refrigeration system is generally satisfactory (*30*), but polymerization is possible. Many of the HFOs were initially studied as polymer precursors, and HFO polymers are fluorinated analogs of the common plastic polypropylene. Much of the information on HFO polymerization is anecdotal, although Richter *et al.* report the results of stability testing on HFO-1234yf; polymerization was associated with high temperatures and high pressures (exceeding those typically encountered in refrigeration systems) and also by the presence of air (which is excluded in properly maintained refrigeration systems) (*31*).

The atmospheric breakdown products of the HFOs are receiving increased attention. Any compound with a $-CF_3$ group in the molecule has the potential to react in the atmosphere to form trifluoroacetic acid, or TFA (CF_3CO_2H). TFA readily dissolves in water and is quickly rained out of the atmosphere. The compound is toxic to aquatic life and does not readily degrade. The breakdown of 1 kg of HFO-1234yf yields 1 kg of TFA. This compares with a TFA yield of 23% for HFC-134a and 10% for HFO-1234ze(E) (*32*). Natural sources of TFA exist along with a natural background level in the oceans. Because of the very short atmospheric lifetimes of HFOs, localized deposition of TFA is a concern (*32*, *33*), and accumulation is possible for basins that do not drain to the ocean.

No consensus exists on whether TFA presents a problem. The World Meteorological Organization in its 2018 assessment of ozone depletion stated that "Atmospheric degradation of HFC-1234yf… produces TFA. Potential impacts beyond a few decades of this TFA source could require future evaluation due to the environmental persistence of TFA and uncertainty in future emissions of HFC-1234yf…" (*28*). The 2018 United Nations Environment Programme (UNEP) Technical Options Committee assessment (*34*) states, "Further research into the environmental impact of… TFA and long term accumulation effects from existing and alternative HFC/HFO refrigerants is needed."

"No perfect refrigerant exists, and trade-offs must be balanced among GWP, ODP, toxicity, flammability, stability, energy efficiency, system complexity, refrigerant price, and prospect of long-term availability."

Resurgence of the natural refrigerants

Ammonia, CO_2, and the hydrocarbons all have 0 ODP and very low GWP, making them environmentally attractive possibilities. Collectively, these are referred to as "natural refrigerants" because they occur in nature, although ammonia is a manufactured chemical, the hydrocarbons are refined from petroleum, and CO_2 is a by-product of various industrial processes. From a safety standpoint, water would be the ultimate refrigerant and is also included in this group.

The natural refrigerants have long been used in industrial systems, in part because of the relatively high cost of fluorocarbons. Many consider ammonia to be a superior refrigerant because of its thermodynamic properties, and application is relatively easy in large, low-temperature systems.

Lorentzen sparked interest in CO_2 in a wider range of refrigeration systems (*35*), and CO_2 is being used across a wide range of systems such as supermarkets, ice rinks, heat pump water heaters, data center cooling, automotive air conditioning, and industrial freezers. But to engineer CO_2 systems to be as efficient as simple fluorocarbon systems requires greater complexity, which usually makes them more expensive to construct and maintain.

Hydrocarbons are used for refrigeration in oil refineries. In this case, the entire plant is designed to handle large tonnages of flammable product so that the flammability hazard of a hydrocarbon refrigerant is readily dealt with.

In very small systems, such as domestic refrigerators, isobutane is commonly used because it gives efficient performance at low cost. The quantity required is also sufficiently low to make the flammability hazard feasible to manage. The use of hydrocarbons in refrigerators was kickstarted in 1992 when the environmental group Greenpeace championed the idea of a household refrigerator that uses a hydrocarbon refrigerant and coined the name "Greenfreeze." Greenpeace collected orders for this product before it was even produced and recruited a near-bankrupt manufacturer in the former East Germany to make it. By 2020, every major refrigerator manufacturer had hydrocarbon units in their product range, and these accounted for 75% of global household refrigerator production (*24*). Although only a small fraction of current refrigerators in North America use hydrocarbons, the EPA SNAP regulations list HFC-134a as "unacceptable" effective 1 January 2021 (*36*), and isobutane is a leading replacement candidate.

The middle ground, in terms of system size and complexity—for example, residential central air-conditioning systems—is where the greatest debate on refrigerant choice occurs. Ammonia is not suitable for small systems or those in close proximity to the general population, and the charge of hydrocarbon, usually propane or isobutane, is larger than can be used without incorporating substantial safety measures.

In the United States, the SNAP regulations allow the use of CO_2 in applications that include refrigeration, air conditioning, process cooling, and refrigerated transport. Ammonia is allowed in industrial settings as well as a wider range of applications provided that a secondary coolant loop is used to isolate the ammonia (in a machine room, for example) from the general public. Hydrocarbons are allowed in household refrigerators and retail standalone display cases subject to a maximum charge of 150 g.

New system configurations

In many cases, the successful application of alternative refrigerants requires the adoption of alternative system configurations—especially for CO_2, the physical properties of which are substantially different to those of the fluorocarbons being replaced. Incremental steps are continually being implemented to improve efficiency, and many variations on the basic vapor-compression cycle exist. Our focus is limited to innovative system configurations that allow either (i) substantially reduced refrigerant charge or (ii) use of new classes of refrigerants in a given application.

Several manufacturers and research groups have demonstrated a wide range of refrigeration systems that use ammonia but with greatly reduced charge size compared with those of traditional technologies (*37*, *38*). These systems have mainly been applied to cold storage facilities where ammonia is familiar, but they have also been used in indirect systems where the ammonia remains separate from the building occupants. A reduced refrigerant charge would also mitigate the hazards of hydrocarbon-based systems.

Many reduced-charge designs use microchannel heat exchangers for the condenser and evaporator. The refrigerant flow channels in these designs have dimensions of <1 mm. Heat transfer in such small channels is increased, and the channels often have internal fins or other structures to further enhance heat transfer. Many parallel channels are incorporated into a single extrusion to reduce pressure drops. This contrasts with more traditional designs in which refrigerant flows in plain round tubes with a diameter of ~10 mm. Microchannel heat exchangers also represent a means of increasing the efficiency of refrigeration equipment apart from the benefit of reduced refrigerant charge (*39*).

A variation on the reduced-charge theme is to use a secondary loop in an indirect system. This has long been done with large, central-system chillers for air conditioning, which cool a flow of water that is pumped to remote air handlers (water-to-air heat exchangers) to provide cooling throughout a building. This arrangement both minimizes and contains the refrigerant to a machine room. By avoiding long refrigerant-filled piping runs, leakage of refrigerant is also greatly reduced. This approach can be very effective in supermarkets, replacing leak-prone piping runs of R-404A (GWP_{100} = 3943) with CO_2 loops to display cases (*40*, *41*).

Refrigerants with a relatively high critical temperature—those operating at relatively low pressures—are intrinsically more energy efficient because they avoid some of the thermodynamic inefficiencies encountered when operating close to the critical point. But in most types of systems, low-pressure refrigerants suffer decreased efficiency because of pressure drops in the condenser, evaporator, and piping. Chiller systems often use low-pressure refrigerants and operate with high efficiency because of different compressor technology and heat exchanger designs that minimize pressure drop. High critical temperatures are associated with larger molecules, and there are many more refrigerant possibilities as molecular size increases, simply because more permutations and combinations for arranging atoms are possible. A number of low-pressure, low-GWP fluids with an "A1" safety rating exist, including HFO-1336mzz(Z), HFO-1336mzz(E), HCFO-1233zd(E), and HCFO-1224yd(Z) (*13*). Cogswell and Verma, for example, describe a small commercial air-conditioning system that uses a centrifugal compressor and low-pressure refrigerant (*42*).

Water presents the extreme example of a low-pressure refrigerant, but a substantially different design would be required to use it. Various research groups have studied such systems, and one company is producing 35-kW systems for process cooling (such as for data centers) (*43*). Whether this technology can be economically applied is an open question.

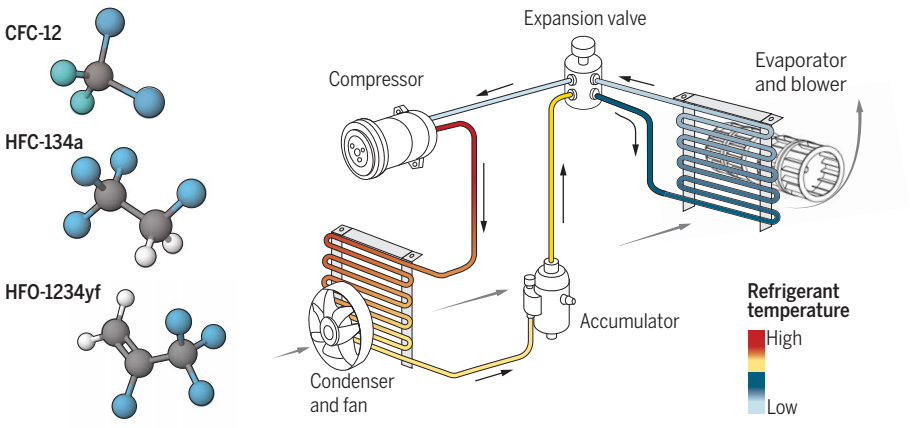


Fig. 3. Progression of refrigerants in different applications.

Trade-offs

The transition from refrigerants with high ODP or GWP to less damaging alternatives has necessitated some trade-offs in capacity, efficiency, operating pressure, and flammability. These have generally been accommodated by the development of more refined component and system designs, but occasionally at the expense of higher production cost or increased system complexity. Taking past examples, the transition from CFC-12 to HFC-134a allowed equipment of similar volumetric capacity, but energy efficiency was decreased by 5 to 10%. This was countered by improved compressor and heat exchanger design so that modern HFC-134a equipment is more efficient than CFC-12 systems of 30 years ago. The transition from HCFC-22 to R-410A enabled smaller compressors to be used for a given duty but increased the operating pressure of the system. The industrial refrigerant R-502 (a blend of a CFC and a HCFC) enabled low temperatures to be achieved with a single-stage compression; substitution with HFC-134a or a blend often required a more complex two-stage compression system to keep the compressor discharge temperature within tolerable limits.

Recently, the lower (class 2L) flammability of some of the HFOs has received considerable attention, and the 2019 revision of the ASHRAE safety standard (*11*) spells out means to mitigate the hazard of 2L refrigerants. Although flammability is inherent to some of the low-GWP refrigerants (most notably, hydrocarbons), it is not inevitable. In the case of the HFOs, a higher substitution of fluorines for hydrogens is required to suppress flammability, so that there are few nonflammable HFOs.

The replacement of R-410A in air-conditioning equipment is a particular challenge. The good energy efficiency of this refrigerant is related to its relatively high operating pressure. This also yields a high volumetric capacity, which is the refrigeration effect per unit of refrigerant vapor volume entering the compressor. A high volumetric capacity is associated with smaller-sized equipment, and current equipment is optimized for the properties of R-410A. There are no nonflammable, single-component refrigerants that are a direct replacement for R-410A. Moving to a low-GWP fluid would mean either (i) a lower volumetric capacity, requiring larger equipment and (perhaps) lower efficiency; (ii) a flammable refrigerant, requiring safety measures; (iii) use of a blend with a moderate value of GWP; or (iv) extensive redesign to match the properties of a different refrigerant (or some combination of these).

No compromise on toxicity is being seriously pursued. With the exception of a few refrigerants for large industrial systems, all of

the new fluids being developed have a toxicity classification of A. The HCFO refrigerants have a nonzero ODP, but these fluids are being developed only for industrial systems.

Summary and discussion

The refrigerants used in vapor-compression equipment have continually evolved in response to changing demands on safety and environmental characteristics. The high GWP of the current generation of refrigerants is compelling the refrigeration industry to consider new fluids. No perfect refrigerant exists, and trade-offs must be balanced among GWP, ODP, toxicity, flammability, stability, energy efficiency, system complexity, refrigerant price, and prospect of long-term availability. Fluorinated olefins (HFOs) have emerged as a new class of refrigerants. The HFOs, often blended with current HFC refrigerants, are the alternative of choice for many applications. The so-called natural refrigerants (including ammonia, CO_2, propane, and isobutane) are seeing renewed interest in a wide range of applications. Some industry sectors have adopted a new path (for example, the use of HFO-1234yf in automotive air-conditioning or isobutane in household refrigerators), but most sectors are still evaluating options. Innovations in system design offer the possibility to expand the choice of refrigerant, although industry is largely staying with conventional technology.

Although we have focused on refrigerant selection, merely swapping a high-GWP refrigerant for one with a low GWP is not the only consideration. Safety, reliability, and energy efficiency must also be retained. The actual goal is to satisfy the objectives of refrigeration (comfort air-conditioning or preservation of food, for example) with minimal impact on the environment, which includes maximizing energy efficiency. In this regard, designing the system to be as energy and resource efficient as practical is paramount. In addition, the equipment must be properly operated (for example, by not over-cooling a space) and maintained to realize maximum energy efficiency. The benefits of careful building or process design to minimize the heat load served by the air-conditioning or refrigeration system has the greatest influence of all. Best of all, there is a resurgence of interest among industry engineers in the development of new systems and refrigerants to meet the energy consumption and environmental needs of tomorrow's society.

REFERENCES AND NOTES

1. G. Constable, B. Somerville, *A Century of Innovation: Twenty Engineering Achievements that Transformed our Lives* (John Henry Press, 2003).
2. D. Coulomb, J.-L. Dupont, V. Morlet, *35th Informatory Note on Refrigeration Technologies. The Impact of the Refrigeration Sector on Climate Change* (International Institute of Refrigeration, 2017).
3. S. Solomon, R. R. Garcia, F. S. Rowland, D. J. Wuebbles, *Nature* **321**, 755–758 (1986).
4. G. Myhre *et al*., in *Climate Change 2013: The Physical Science Basis, Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, 2013).
5. G. J. M. Velders, S. O. Andersen, J. S. Daniel, D. W. Fahey, M. McFarland, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 4814–4819 (2007).
6. G. J. M. Velders, D. W. Fahey, J. S. Daniel, S. O. Andersen, M. McFarland, *Atmos. Environ.* **123**, 200–209 (2015).
7. G. J. M. Velders, D. W. Fahey, J. S. Daniel, M. McFarland, S. O. Andersen, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10949–10954 (2009).
8. M. O. McLinden, D. A. Didion, *ASHRAE J.* **29**, 32–42 (December, 71987).
9. T. Midgley Jr., *Ind. Eng. Chem.* **29**, 241–244 (1937).
10. S. Kujak, K. Schultz, *Sci. Tech. Built Environ.* **22**, 1226–1237 (2016).
11. ASHRAE, "ANSI/ASHRAE Standard 15—2019 Safety Standard for Refrigeration Systems," (ASHRAE, 2019).
12. International Organization for Standardization (ISO), "ISO 5149:2014 Refrigerating systems and heat pumps—Safety and environmental requirements Parts 1 to 4" (2014); www.iso.org/standard/54979.html.
13. ASHRAE, "ANSI/ASHRAE Standard 34-2019 Designation and Safety Classification of Refrigerants," (ASHRAE, 2019).
14. ISO, "ISO 817:2014 Refrigerants—Designation and safety classification" (2014); www.iso.org/standard/52433.html.
15. J. M. Calm, *Int. J. Refrig.* **31**, 1123–1133 (2008).
16. J. He, T. M. Tritt, *Science* **357**, eaak9997 (2017).
17. X. Moya, N. D. Mathur, *Science* **370**, 797–803 (2020).
18. J. S. Brown, P. A. Domanski, *Appl. Therm. Eng.* **64**, 252–262 (2014).
19. A. Kazakov, M. O. McLinden, M. Frenkel, *Ind. Eng. Chem. Res.* **51**, 12537–12548 (2012).
20. P. A. Domanski, R. Brignoli, J. S. Brown, A. F. Kazakov, M. O. McLinden, *Int. J. Refrig.* **84**, 198–209 (2017).
21. M. O. McLinden, J. S. Brown, R. Brignoli, A. F. Kazakov, P. A. Domanski, *Nat. Commun.* **8**, 14476 (2017).
22. National Oceanic and Atmospheric Administration, The 2019 ozone hole is the smallest ever recorded (2019); www.noaa.gov/news/2019-ozone-hole-is-smallest-ever-recorded.
23. J. S. Brown, *HVAC R Res.* **19**, 693–704 (2013).
24. C. Booten, S. Nicholson, M. Mann, O. Abelaziz, "Refrigerants: Market Trends and Supply Chain Assessment," National Renewable Energy Laboratory technical report NREL/TP-5500-70207 (2020).
25. Y. Kayukawa, N. Sakoda, R. Akasaka, *Trans. JSRAE* **37**, 1–44 (2020).
26. M. Ito, Z. Zhang, C. Dang, E. Hihara, in 25th IIR International Congress of Refrigeration, paper 0822 (Montreal, 2019).
27. M. Hashimoto *et al*., *Sci. Tech. Built. Environ.* **25**, 776–783 (2019).
28. World Meteorological Organization (WMO), Scientific assessment of ozone depletion: 2018, Global Ozone Research and Monitoring Project–Report no. 58. (WMO, Geneva, Switzerland, 2018).
29. K. Schultz, paper presented at the 25th International Congress of Refrigeration, paper 0144 (Montreal, 2019).
30. S. Kujak, E. Sorenson, *ASHRAE J.* **60**, 28–35 (2018).
31. M. Richter, M. O. McLinden, E. W. Lemmon, *J. Chem. Eng. Data* **56**, 3254–3264 (2011).
32. K. R. Solomon *et al*., *J. Toxicol. Environ. Health B Crit. Rev.* **19**, 289–304 (2016).
33. J. Kazil *et al*., *J. Geophys. Res. Atmos.* **119**, 14059–14079 (2014).
34. UNEP, Refrigeration, Air Conditioning and Heat Pumps Technical Options Committee, 2018 Assessment Report (United Nations Environment Programme, 2019).
35. G. Lorentzen, *Int. J. Refrig.* **17**, 292–301 (1994).
36. 40 CFR Part 82. Protection of Stratospheric Ozone: New Listings of Substitutes (2016); www.govinfo.gov/content/pkg/FR-2016-12-01/pdf/2016-25167.pdf.
37. A. B. Pearson, D. Hamilton, paper presented at the 11th IIR Gustav Lorentzen Conference on Natural Refrigerants, Hangzhou, China, 2014, 2014.
38. P. Hrnjak, paper presented at the 8th IIR Conference: Ammonia and CO_2 Refrigeration Technologies, Ohrid, 2019.
39. M.-H. Kim, S. Y. Lee, S. S. Mehendale, R. L. Webb, in *Advances in Heat Transfer*, vol 37. (Academic Press, 2003).
40. A. Pearson, A. Campbell, *ASHRAE J.* **52**, 24–28 (2010).
41. R. Lazzarin, Y. Ge, C. Zilio, *37th Informatory Note on Refrigeration Technologies. Advancements in Supermarket Refrigeration* (International Institute of Refrigeration, 2018).
42. F. Cogswell, P. Verma, *High-Efficiency Low Global Warming Potential Compressor; final report to U.S. Department of Energy on contract DE-EE0007039* (United Technologies Research Center, 2018).
43. Efficient Energy GmbH, eChiller—The refrigeration system with the safe and natural refrigerant water (R718) (2020); https://efficient-energy.de/en/the-most-efficient-chiller.

10.1126/science.abe3692

REVIEW

Caloric materials for cooling and heating

X. Moya and N. D. Mathur

Magnetically driven thermal changes in magnetocaloric materials have, for several decades, been exploited to pump heat near room temperature. By contrast, their electrocaloric and mechanocaloric counterparts have only been intensively studied and exploited for little more than a decade. These different caloric strands have recently been unified to yield a single field of research that could help combat climate change by generating better heat pumps for both cooling and heating. Here we outline the timeliness of the present activity and discuss recent advances in caloric measurements, materials, and prototypes.

The evolution of research into magnetocaloric (MC), electrocaloric (EC), and mechanocaloric (mC) materials has been asynchronous and saltatory ever since the first salvos were metaphorically fired during the Battle of Trafalgar, but only in relatively recent times have these magnetically, electrically, and mechanically driven thermal changes been exploited in prototype heat pumps (*1*). All three strands have been unified even more recently under the banner of caloric materials, with mC materials representing the nexus of elastocaloric (eC) and barocaloric (BC) materials, driven by changes of uniaxial stress and pressure, respectively. Away from absolute zero, caloric effects are commonly much enhanced by driving a phase transition that may be first-order or continuous. Field-driven phase transitions can only be defined along the first-order phase boundaries that terminate at some critical point (Fig. 1), and substantial caloric effects in the untransformed and transformed phases can augment the large caloric effects that arise from the transition itself (Fig. 2).

Caloric effects are typically parameterized as changes of temperature ΔT in the adiabatic limit, or changes of entropy ΔS in the isothermal limit. The isothermal heat $Q = T\Delta S$ is reported less often, even though this directly measurable quantity sets the scale for the heat pumped in cooling and heating cycles. Caloric effects can be repeatably exploited in consecutive cycles, even if they are not thermodynamically reversible, as long as they are nominally reversible. Here we define a caloric effect to be nominally reversible if the application and removal of a field causes the material to return to a state that is macroscopically if not microscopically indistinguishable from the starting state, whether or not there is any field hysteresis, such that the thermal changes on field application and field removal are equal in magnitude.

The adiabatic temperature change that can be achieved in a caloric working body need not limit the temperature difference against which heat is pumped. By using regeneration, the caloric working body can establish a relatively large temperature differential along a displaceable heat-transfer medium, which is often a fluid (passive regeneration) (*2*), or along both itself and such a medium (active regeneration) (*3*, *4*). By contrast, in the well-established process of vapor compression, the fluid working body is itself responsible for heat transfer. Vapor compression continues to form the basis of many cooling systems as well as a growing number of heating systems that are starting to replace inefficient gas-fired boilers. However, vapor-compression fluids remain problematic because global warming potential can only be reduced at the cost of flammability or performance (*5*). Moreover, vapor compression can be slow to start up, noisy, inefficient under variable-load conditions, and hard to miniaturize.

Caloric prototypes have been proposed for both cooling and heating (*1*, *6*) but have yet to reach the performance levels achieved by vapor compression. Caloric materials could therefore first find use supporting vapor compression in hybrid devices, which would mirror the hybrid devices based on thermoelectrics and vapor compression (*7*). Alternatively, caloric materials could improve the efficiency of thermoelectrics in all-solid-state hybrid devices (*8*). Unlike standard implementations of vapor compression, some caloric prototypes now recover some of the recoverable work that is done to drive caloric effects (*9*). This recovery is automatic if the caloric working body circulates through a region of field, as seen for quasi-continuous MC discs (*10*), and as possible for EC fluids (*11*). Otherwise, work can be recovered by using two caloric bodies that help drive each other in antiphase, as demonstrated for eC (*12*, *13*) and EC (*14*, *15*) prototypes.

The recent activity on calorics is reflected in a growing number of publications (Fig. 3) on materials and prototypes (Fig. 4) and in ever more meetings and networks, most notably the caloric sessions at meetings of the Materials Research Society and the inclusion of all types of caloric effect at the international conference on caloric cooling (THERMAG). The recent activity is timely given that caloric materials are included in a 2017 U.S. Department of Energy report on thermal management in commercial buildings (*16*) and in a U.K. roadmap that addresses the 2050 global challenge of decarbonization (*17*). Nevertheless, the focus need not be wholly applied, and it is important not to lose sight of the science that has the capacity to

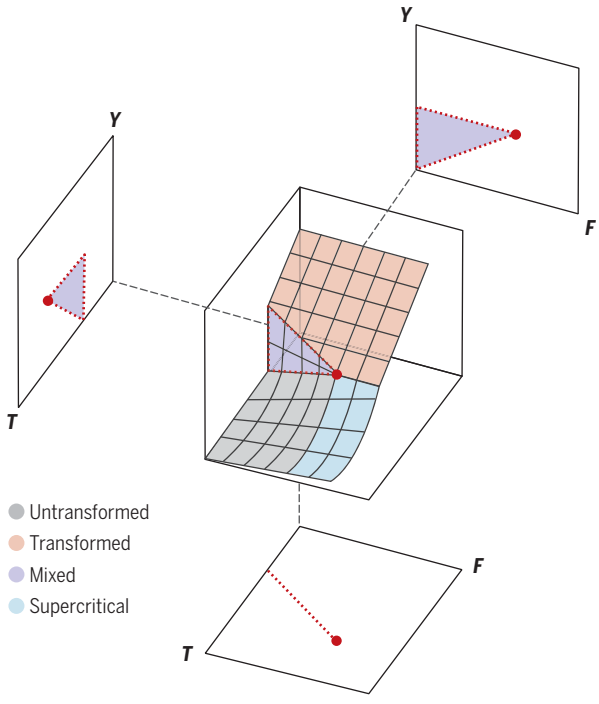


Fig. 1. Caloric effects near a first-order phase transition. The 3D phase diagram shows order parameter *Y* versus field *F* and temperature *T* for conventional caloric effects. The generalized field *F* is magnetic field (MC), electric field (EC), uniaxial stress (eC), or pressure (BC). The generalized order parameter *Y* is magnetization (MC), polarization (EC), uniaxial strain (eC), or inverse volume (BC). We assume that the first-order phase transition is experimentally sharp at zero field and anhysteretic. Black grid lines denote contours of constant *F* and *T*; red dots denote the critical point. The mixed-phase region (purple) is bounded by red dotted lines and projected onto three two-dimensional phase diagrams. Caloric effects arise from both the field-driven phase transition and the untransformed and transformed phases themselves (Fig. 2).

Department of Materials Science, University of Cambridge, Cambridge, UK.
Email: xm212@cam.ac.uk (X.M.); ndm12@cam.ac.uk (N.D.M.)

transcend any technological impact. As with our 2014 review (*1*), readers are encouraged to search for inspiration in sections featuring caloric effects on which they do not currently work. As always, heat and work will alas be used to describe energies as if they were not in transit.

Caloric measurements

It is important to verify reversibility when measuring any type of caloric effect. Caloric effects associated with continuous transitions are likely to be thermodynamically reversible, whereas caloric effects associated with first-order transitions show a field hysteresis that, whether large or small, is inextricably linked with a corresponding thermal hysteresis (Fig. 5A). Although this field hysteresis necessarily precludes thermodynamic reversibility, it is nevertheless possible to drive nominally reversible caloric effects in full by avoiding the zero-field thermal hysteresis and operating at starting temperatures that lie at and above $T_{h2}(0)$ (Fig. 5, A and B, orange and green arrows). Alternatively, after arriving at the zero-field heating branch in $T_{h1}(0) < T < T_{h2}(0)$ by either heating or completing one field cycle, the transition can be partially or fully completed in a nominally reversible manner using fields that are smaller than the fields required to drive the transition in full at and above $T_{h2}(0)$. Note that caloric effects in the untransformed and transformed phases themselves (not shown in Fig. 5A) would be liable to modify the above analysis to some extent.

Indirect, direct, quasi-direct, and quasi-indirect measurements are described in (*1*) and summarized in Table 1. Indirect methods dominate the literature because they are circumstantially easy. They are based on thermodynamic analysis of order parameter data that are typically obtained by measuring isothermal hysteresis loops during a unidirectional temperature sweep, but other measurement protocols are also possible (Table 1). The thermodynamic analysis assumes that the order parameter is a single-valued function of field and temperature, which is reasonable if any hysteresis in field and temperature is small. Unipolar loops reassuringly mimic the field sweeps that are used in direct measurements, and they provide a direct test of both reversibility and nominal reversibility, such that they should be preferred to bipolar loops (whose inner branches cannot be used if ferroic domains are present). For materials that display conventional caloric effects at first-order transitions, loops should be measured on heating to ensure that caloric effects identified near the transition are nominally reversible. By contrast, cooling data are required for inverse caloric effects (*1*). Widespread failure to thus sweep temperature in the correct direction can render literature comparisons problematic.

The indirect nature of indirect measurements should not be used as an excuse to employ subsequent thermodynamic analysis without confirming that transitions are exploited while measuring loops. For continuous transitions, a field-driven increase of order parameter should be observed. For first-order transitions, the order parameter should display a hysteretic step at start and finish fields (temperatures) that vary with temperature (field) (Fig. 5A). This variation creates sloping phase boundaries, or a single boundary if the transition is identified more crudely by way of a single transition temperature T_0, and gradients are given by the Clausius-Clapeyron equation (*1*). To drive transitions with as small a field as possible, it is desirable for the phase boundaries to slope with a shallow gradient and for transitions to be narrow (narrowness implies limited interaction between untransformed and transformed regions). Impractically large fields may be required to traverse the steep boundary between phases that differ greatly in entropy but little in order parameter, whereas infinite steepness would wholly preclude any caloric effect. One recent improvement to the indirect method is to identify values of ΔT (values of ΔS) by tracing adiabatic (isothermal) contours on high-resolution entropy (temperature) maps that are constructed from isothermal (adiabatic) hysteresis loops, most notably because there is no need to assume a constant heat capacity (*18*, *19*).

Direct measurements of temperature change are important for unambiguous verification of the large ΔT values that drive the flow of heat. In MC materials, direct measurements of temperature change typically comprise wholly adiabatic excursions either above or below the starting temperature, thus providing a direct test of nominal reversibility. In EC and mC materials, direct measurements of temperature change typically permit the comparison of highly adiabatic heating and cooling legs that are each followed by a slow return to the starting temperature (unless Joule heating compromises electrically leaky EC materials). By considering this cycle on axes of entropy and temperature, it can be appreciated that the highly adiabatic legs will differ in length if the isofield legs that they separate are nonlinear. Field-on and field-off ΔT values can thus show an intrinsic difference in magnitude, such that they need not differ purely as a consequence of any irreversibility (*18*, *20*, *21*).

Direct measurements of heat based on variable-field calorimetry provide a direct test of reversibility, and for EC materials it is common for the changes of field to be fast and highly adiabatic, as when measuring temperature change, rather than slow and isothermal as expected. This tends to be reasonable when quantifying isothermal EC effects near room temperature, but errors can arise if the calibration and measurement runs are not performed on the same time scale, most notably when calibrating with steady-state Joule heat from a resistor and simultaneously measuring samples via thermal barriers. Thermal barriers also diffuse the flow of heat to the extent that the signal from a first-order transition can become lost in the baseline. Examples of thermal barriers include the electrically insulating glue that is sometimes used instead of Kapton tape to isolate EC samples, and the pressure media required to drive BC effects.

Quasi-direct measurements of heat provide an indirect test of nominal reversibility if the fixed-field temperature sweeps are performed on both heating and cooling. Quasi-direct measurements of heat typically require baselines to be modified by the magnitude and sign of any non-negligible caloric effects in the untransformed and transformed phases themselves (*22*). The quasi-direct method is attractive when direct measurements of heat are challenging, for example, in the case of eC effects associated with large deformations. The quasi-direct method is also attractive because a viable dataset can be efficiently collected by sweeping the temperature at ever-larger fields prior to any breakdown. Breakdown most commonly afflicts EC and eC materials, such that choosing the maximum field is akin to gambling at a casino, where raising the stakes invariably carries the risk of going bust.

Quasi-indirect measurements of temperature change were demonstrated in 1997 using an MC material (*23*) and named as such in recent

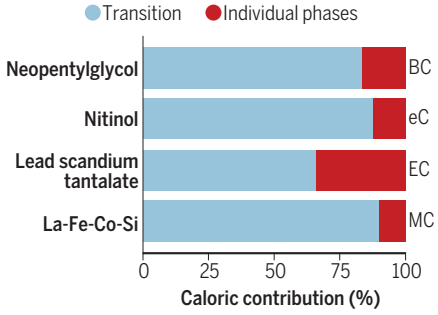


Fig. 2. Caloric contributions in well-known materials. For conventional caloric effects in La-Fe-Co-Si (MC, 5 T), lead scandium tantalate (EC, 29 V μm^{-1}), nitinol (eC, 600 MPa), and neopentylglycol (BC, 0.25 GPa), the thermal change due to the field-driven phase transition (blue) is augmented by thermal changes in the untransformed and transformed phases themselves (red). Thermal changes were estimated at specific temperatures that lie above the relevant transition temperatures, using ΔS for lead scandium tantalate (*18*) and neopentylglycol (*136*) and ΔT for La-Fe-Co-Si (*49*) and nitinol (*101*). The contribution of the transformed phases could be increased by using a larger field. If a transition and a transformed phase yield caloric effects of opposite sign (*22*), then the transition should be driven using the smallest possible field.

work on an EC material (*24*). When comparing adiabats and isotherms that describe field-driven changes of order parameter, values of ΔT are identified for an adiabat by identifying the temperature difference between two isotherms that it spans, and it can be helpful to render many nearby isotherms on a periodic color scale for visual interpretation of dense data (*24*). If the changes in the field-driven order parameter are found to be nominally reversible, then the nominal reversibility of the corresponding caloric effects can also be inferred. Quasi-indirect measurements avoid the heavy processing and possible artifacts that can corrupt the indirect method, and quasi-indirect measurements obviate the challenge of measuring small samples via contact thermometry because the sample itself indirectly functions as the thermometer.

A key advance in recent years has been the addition of infrared imaging to the caloric toolkit (*1*), thus permitting real-space visualization of temperature change and heat flow in monolithic materials (*21*, *25*–*27*), multilayer capacitors (MLCs) based on EC materials (*18*, *28*, *29*), and prototypes (*14*, *30*–*32*). However, care must be taken to ensure good calibration and account for unwanted reflections (*8*, *18*). In the future, infrared imaging could be routinely used to measure low-thermal-mass samples, such as thin films without substrates.

Magnetocalorics

The recent progress on monolithic and composite materials belies the maturity of the field. Developments include large isochoric MC effects without cracking in MnFe(P,Si,B) (*33*), an annealing-induced 50 K variation of operating temperature in $Fe_{48}Rh_{52}$ (*34*), and considerable activity on molecular magnets (*35*) and amorphous metals (*36*). Polymer-MC composites can demonstrate good mechanical integrity after 90,000 cycles (*37*), and their optimization can involve a multiparameter space that includes particle diameter (*38*). By contrast, metal-MC composites, such as those where MC particles are coated with a low-melting eutectic alloy (*39*) or hot-pressed with Sn (*40*), can show superior mechanical, heat-transfer, and corrosion properties. More generally, corrosion can be avoided by applying a surface coating to protect against the water commonly used for heat transfer (*41*), albeit at the cost of extra thermal resistance. Meanwhile, academic work on MC films continues to progress (*42*).

In Heusler alloys, nominal reversibility has been achieved by partially driving a hysteretic first-order phase transition from the mixed phase to ensure that nucleation is not a limiting factor (*43*, *44*), and the magnitude of these MC effects can depend on how quickly the measurement temperature is reached (*44*). Moreover, hysteresis can be minimized by reducing microstructural complexity, for example, by avoiding precipitation and segregation (*45*) and by maximizing the structural compatibility of the two phases (*46*). It has also been shown that ionic bombardment can create defects that assist nucleation (*47*) and that limited heat exchange can increase hysteresis (*48*).

Metrological developments include 1- to 10-Hz measurements for compatibility with prototypes (*49*–*51*), faster measurements for the infrared detection of 1.4-μm-thick films (*52*), and 62-T pulsed-field measurements that reveal MC effects in Gd to be enhanced above the Curie temperature (*53*). MC measurements have also been used to establish an unambiguous method for determining the order of magnetic phase transitions (*54*). Unfortunately, large MC effects are sometimes identified erroneously, for example, when the magnetization rotates without other changes and when field-driven transition temperatures shift too little for transitions to be magnetically driven in practice. Fortunately, the decomposition of MC effects into electronic, magnetic, and lattice contributions is now less well entrenched in strongly coupled materials (*55*).

The development of MC prototypes continues apace, primarily using working bodies that comprise beds of commercial-grade Gd spheres (*56*). Rotating active magnetic regenerators (AMRs) that exploit automatic energy recovery (*9*) have been used in a range of systems, for example, in a prototype based on 2.8 kg of Gd. The extremes of the resulting load line permit a cooling power of 1010 W or a temperature span of 25.4 K (*3*). By contrast, a prototype based on 1.52 kg of variable-composition $La(Fe,Si)_{13}H$ displays 3042 W at zero temperature span, or 2502 W at a temperature span of 11 K (*57*). Future improvements could arise from detailed AMR modeling, which has already revealed the value of combining materials (*58*–*60*) and varying bed geometry (*61*).

Electrocalorics

The development of monolithic EC materials in recent years has proceeded in parallel with more complex implementations that include MLCs. Ceramic and polymer materials continue to dominate, but lead-free ceramics have attracted growing interest, and ceramic nanoparticles have been introduced into polymer

Table 1. Methods of caloric measurement. TVoM, target variable of measurement. The generalized field F is magnetic field (MC), electric field (EC), uniaxial stress (eC), or pressure (BC). The order parameter Y is magnetization (MC), polarization (EC), uniaxial strain (eC), or (unlike Fig. 1) negative volume (BC). Start temperature is denoted T_s, time is denoted t. Any values of ΔT, ΔS, and Q should be interconverted using entropy maps (*18*, *19*), as the traditional interconversion ($c|\Delta T| \sim T|\Delta S| = |Q|$) requires some effective value of c and incorrectly assumes the same driving field for ΔT and ΔS (Fig. 5). The driving field for ΔT is therefore often underestimated.

Method	Excursion	TVoM	Caloric analysis	Use
Indirect	Isothermal F sweeps at many T	Y	$\Delta S(F) = \int(\partial Y/\partial T)_S dF$ at each T	common
	Isofield T sweeps at many F	Y	$\Delta S(F) = \int(\partial Y/\partial T)_S dF$ at each T	rare
	Adiabatic* F sweeps at many T_s	Y	$\Delta T(F) = -\int(\partial Y/\partial S)_T dF$ at each T_s†	rare
Direct	Adiabatic* F sweep at any T_s	ΔT	N/A	common
	Isothermal F sweep at any T	Q‡	N/A	rare
Quasi-direct	Isofield T sweeps at many F	Q‡	Subtract $S(T) = \int dQ/T$ at each F to obtain $\Delta S(F)$ at each T	used
Quasi-indirect	Isothermal F sweeps at many T + Adiabatic* F sweep at any T_s	Y + Y	Identify ΔT for each adiabatic $Y(F)$ plot by comparing with isothermal $Y(F)$ plots	rare

*Conditions would ideally be isentropic to avoid all dissipative heating (*8*). †Convert values of T_s to values of S via zero-field heat capacity $c(T)$ (*19*). ‡Record dQ/dt and evaluate $Q = \int(dQ/dt)/|dF/dt|dF$ (direct) or $Q = \int(dQ/dt)/|dT/dt|dT$ (quasi-direct) after typically subtracting the baseline from the integrand, thus most readily capturing discontinuous thermal changes.

films. Preliminary work on other types of material includes the study of liquid crystals (*11*) for automatic energy recovery in prototypes (*9*), ferroelectric plastic crystals (*62*), ferrielectric ammonium sulfate (*63*), and organic-inorganic hybrid ferroelectrics (*64*). Large electric fields are required to drive the large entropy changes in these materials because the concomitant changes of electrical polarization are small, which means that avoiding electrical leakage will be key.

Lead-based ceramics continue to be a relatively popular vehicle for the growing interest in inverse EC effects, which is the correct term for so-called negative EC effects given that ΔT and ΔS take opposite signs to each other for a given sign of field change. Inverse EC effects are commonly inferred via the indirect method if the isofield polarization increases with increasing temperature, for example, at an antiferroelectric-ferroelectric transition (*65*). However, the indirect method is not a good a priori test for EC effects at an order-order transition, because the reorientation of local dipoles could be nominally non-entropic, just like the aforementioned magnetization rotations. Nevertheless, 3.5 K inverse EC effects have now been directly measured at an antiferroelectric-ferroelectric transition using 42 kV cm^{-1} (*66*).

Lead-free ceramics such as $Ba(Hf,Ti)O_3$ (*67*) and $Ba(Zr,Ti)O_3$ (*68*) are reported to show few-kelvin changes near the critical point, but further investigation is warranted to obtain direct measurements of temperature change and identify which three phases meet according to the Gibbs phase rule. The lead-free ceramic $K_{0.5}Na_{0.5}NbO_3$-$SrTiO_3$ (*69*) displays similar changes of temperature in a broad temperature range because it exploits relaxor behavior, and the lead-free Aurivillius-phase relaxor $SrBi_{1.85}Pr_{0.15}(Nb_{0.2}Ta_{0.8})_2O_9$ (*70*) is interesting because it avoids fatigue like its $SrBi_2Ta_2O_9$ (SBT) parent, which was developed for ferroelectric memories (*71*).

The study of lead-free ceramic films has been extended to include thick layers (*72*), bilayers (*73*), and compositionally graded layers (*74*), leading to optimistic values of EC temperature change that were obtained using calorimetry in the presence of substrates. However, the nonadiabatic conditions imposed by substrates do not preclude the direct measurement of EC temperature change in films, as shown using scanning thermal microscopy (*75*) and high-frequency thin-film resistance thermometry (*76*). Free-standing polymer films have been used to make direct EC measurements of temperature change (*77*), and similar measurements would be required to confirm the potentially large EC effects that have been reported for polymer films in which ceramic nanoparticles are intended to increase crystallization and provide an internal biasing field (*78*, *79*). The study of polymer films has been further broadened to include a composite in which the EC polymer was grown as vertical nanowires in a 100-μm-thick holey template of anodic aluminum oxide (*80*), thereby improving two limiting factors, namely crystallization and thermal transport.

Bespoke MLCs based on well-known EC ceramics display electrically driven changes

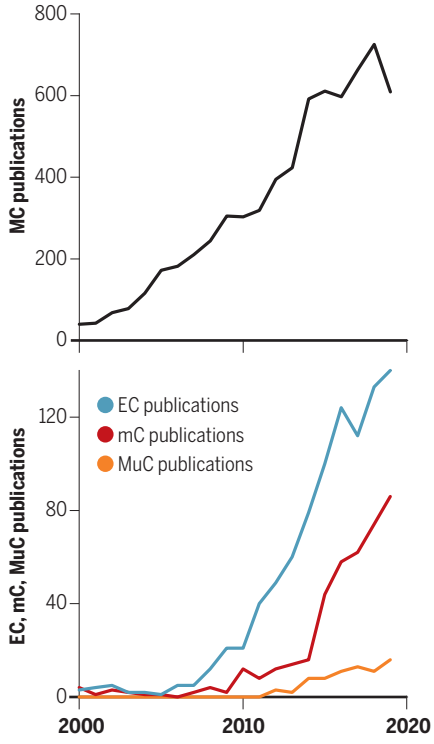


Fig. 3. Caloric publications in the 21st century. MC research remains dominant, EC research and mC research continue to rise following the onset of increased activity in the latter half of the first decade of this century, and the decade now ending has witnessed the emergence of multicaloric (MuC) research involving more than one type of caloric effect. All data are subject to errors associated with search terminology.

of temperature that can readily be measured directly (*18*, *29*, *81*). However, the temperature change at MLC centers should typically be reduced to account for at least some degree of thermalization between active and inactive areas (*18*), not increased by assuming complete internal thermalization prior to measurement (*29*). With or without internal thermalization, MLCs of highly ordered $PbSc_{0.5}Ta_{0.5}O_3$ (PST) display large EC temperature changes that peak at 5.5 K (*18*), thus exceeding the 2.5 K peak for Gd working bodies addressed by permanent magnets in MC prototypes (*56*). Good sample quality was responsible for increasing the breakdown field enough to drive the sharp first-order transition supercritically, thus permitting the EC temperature changes to remain at several kelvin over a wide range of starting temperatures. Given that breakdown in MLCs occurs at electrode edges, it could be desirable to reduce field concentration by inserting layers that help spread charge, as demonstrated with dielectric actuators (*82*).

The performance that EC materials could achieve in operando is now receiving some attention. Detailed material maps permit the construction of accurately parameterized cooling cycles in which balanced regeneration is achieved by varying the finite field applied during regenerator transit (*19*). Separately, it has been observed that cooling power can be optimized by controlling the time-dependent electric-field profile (*83*). On the subject of fatigue, EC effects in the relaxor ferroelectric $Pb(Mg_{1/3}Nb_{2/3})O_3$ were unchanged after 10^6 sufficiently gentle cycles (*84*) but were decimated after 70,000 aggressive cycles, because of enhanced grain boundary conductivity (*85*). By contrast, avoiding degradation in $Ba(Zr_{0.2}Ti_{0.8})O_3$ was accomplished by reversing the electric-field polarity every 10^5 cycles, perhaps reversing oxygen vacancy migration (*86*).

EC prototypes have been under intense development in recent years. For example, solid-to-solid regeneration was achieved using two juxtaposed rings of $BaTiO_3$-based MLCs that rotate in opposite directions, yielding a temperature span of 2 K (*4*). Prototypes with a fluid regenerator (*87*) were substantially improved when $BaTiO_3$-based MLCs were replaced with the aforementioned MLCs of PST, resulting in a temperature span of 13 K or a cooling power inferred to be 1.22 W (*88*). The equivalent replacement in prototypes with mechanical heat switching (*89*) resulted in a temperature span of 5.2 K or a cooling power of 85 mW, and there is scope to optimize the heat switching for improved performance (*90*). These temperature lifts compare well with the value of 12.7 K for a 1995 prototype whose much heavier EC working body comprised 400 g of doped PST plates (*91*). In the future, it would be desirable to increase total MLC mass relative to the total mass of an MLC-based prototype. Alternatively, a relatively low-mass device can be constructed by electrostatically actuating an EC polymer bilayer between sink and load (*92*), just as $BaTiO_3$-based MLCs were translated between sink and load when demonstrating energy recovery (*14*). Stacking four bilayer devices yielded a temperature span of 8.7 K or a cooling power of 906 mW, and energy recovery was successfully implemented (*15*).

Elastocalorics

This area is playing fast catch-up with its better-established MC and EC counterparts, resulting in rapid materials development and the first prototypes. Materials primarily

divide into alloys that display diffusionless (martensitic) transitions (*1*) and polymers that either display crystallization transitions (*93*) or behave like entropic springs (*94*). In addition, small eC effects have been reported in liquid-crystal elastomers (*11*). Large eC effects in ceramics could in the future be realized given that epitaxially grown ceramic membranes can reach 8% uniaxial tensile strain without fracture (*95*). Experimentally, a given study typically tests materials under tension or compression but rarely both (*96*), and mechanical training is typically required to achieve repeatable performance, at least in alloys.

Single-crystal alloys such as $Ni_{50}Fe_{19}Ga_{27}Co_4$ (*97*) show highly anisotropic eC effects, but other single-crystal alloys show much less anisotropy (*98*). The anisotropy may arise because the orientation of the applied stress affects the relative contribution of the phase transition with respect to the elastic heat of a given phase. This elastic heat can be so large that it dominates eC effects arising from the transition (*99*), as previously seen for BC materials (*1*). If a relatively large elastic heat accompanies a mechanically driven transition, then the two caloric components should be driven additively by using a driving stress of the correct sign, thus exploiting a degree of freedom that cannot be meaningfully exploited when driving the other types of caloric effect. Compressive stress would be liable to result in the desired addition for a typical material with positive thermal expansion and thus conventional elastic heat, which is fortuitous given that the propensity for fatigue and failure is greater under tension than it is under compression.

Progress on polycrystalline alloys is primarily centered on the prototypical shape-memory material Ni-Ti known as nitinol (*100*). Rubberlike reversible deformations of up to ~1% strain can be achieved when nanocrystalline martensitic twins undergo reorientation (*101*), and precipitates of Ti_3Ni_4 produce internal stress that inverts the sign of eC effects (*102*). An alloy of Ni-Ti-Mn-B can display a 31.5 K adiabatic temperature change (*103*), which compares well with the value of 28 K (*104*) or higher (*105*) for nitinol. Separately, alloy synthesis by additive manufacturing should prove useful in the construction of bespoke regenerators (*106*).

The pre-elongation of natural rubber, an eC polymer with a mechanically driven crystallization transition, both increases eC strength and reduces the large required driving strains that represent a challenge for applications (*107*). By contrast, the terpolymer poly(vinylidene fluoride-trifluoroethylene-chlorotrifluoroethylene) [P(VDF-TrFE-CTFE)], which is similar to the well-known EC material P(VDF-TrFE-CFE), shows relatively small eC effects (*94*). This is because the coiling and uncoiling of the polymer chains (entropic spring) involve a configurational entropy change that is smaller than the entropy change associated with polymer chain interactions during crystallization.

Fatigue in natural rubber is best avoided by not entering the amorphous phase at low strain, permitting an endurance of 1.7×10^5 cycles, which is comparable with the endurance of alloys (*108*). Fatigue in alloys can be avoided by driving eC effects that arise from elastic heat and no transition (*109*), but transitions are typically key to large eC effects, and so fatigue in alloys is well studied. Fatigue in alloys can be mitigated by creating a nanocomposite that minimizes hysteresis (*110*), introducing a ductile grain boundary phase (*111*, *112*), creating fine textured grains via hot extrusion (*113*), creating nanoscale grains by cold rolling and then high-temperature aging (*114*), selecting for good structural compatibility between austenitic and martensitic phases (*115*), judicious doping (*116*), and cycling in the middle of the transformation plateau (*117*). In single crystals—and, by implication, textured polycrystals—judicious choice of the driving stress direction can mitigate fatigue (*118*). Separately, surface polishing can reduce the propensity for breakdown (*119*).

Following the first eC prototype based on Ni-Ti rods (*120*), a number of prototypes have used Ni-Ti–based foils and films, whose large surface-to-volume ratios permit good heat exchange. For example, a temperature span of 13 K or a cooling power of 120 mW was achieved by translating a 20-μm-thick foil to alternately make conformal contact with a flat source and a convex sink, thus yielding a device whose antiphase operation with respect to another such device permitted energy recovery (*13*). By contrast, the linear deformation of order-of-magnitude–thicker films in a regenerative heat pump yielded a temperature span of 15.3 K and a heating power of 4.64 W (*31*). A similar temperature span was achieved by flowing water through tubes containing Ni-Ti–based eC wires, and three such wires in series produced a temperature span of 28 K or a cooling power of 0.23 W (*121*). The one-shot twisting of similar wires has been used to cool water in part of a wider study that included caloric effects in coiled and supercoiled polymers (*122*), thus reminding that

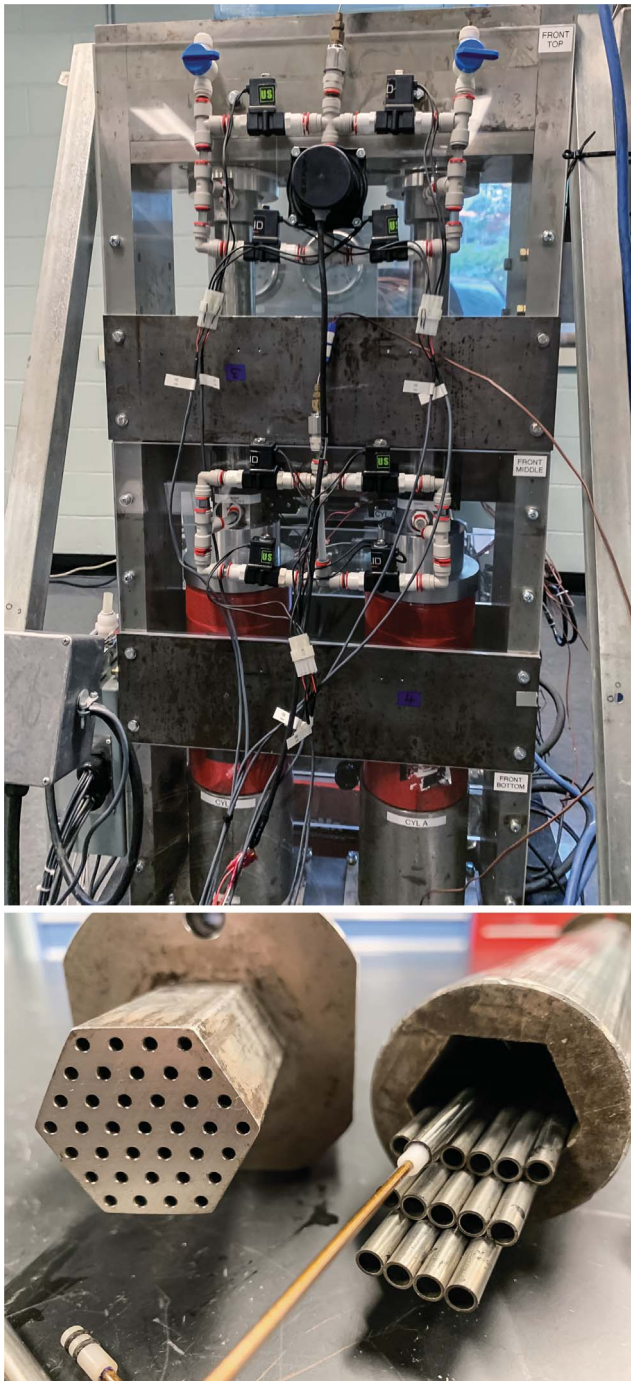


Fig. 4. A recent caloric prototype. The prototype shown here (top) pumps heat using elastocaloric tubes (bottom) in which changes of strain drive changes of temperature.

PHOTOS: ICHIRO TAKEUCHI, UNIVERSITY OF MARYLAND

the stress used to drive eC effects in prototypes need not be wholly linear, as when bending foils (*13*). The use of nonlinear stress inspires us to suggest that eC effects can be defined more broadly to include all forms of deviatoric stress, thus avoiding a taxonomical fragmentation that is automatically avoided when using the mC epithet.

Barocalorics

BC materials were first studied just before the start of the present century (*1*) and have been identified in many types of material. During the past few years, BC effects have been experimentally identified in more magnetic alloys (*123*–*126*), a ceramic ferroelectric (*127*), an ionic salt (*128*), inorganic salts that are ferrielectric (*22*) and ferroelectric (*129*), a fluoride (*130*), organic-inorganic hybrid perovskites (*131*, *132*) and molecular crystals (*133*), and rubber (*134*, *135*). Recently identified BC effects in plastic crystals of neopentylglycol are so large that the entropy change of $|\Delta S| = 510\ \mathrm{J\ K^{-1}\ kg^{-1}}$ is similar to the values obtained when exploiting hydrofluorocarbons in vapor-compression refrigeration (*136*).

The pressure-driven heat of a BC transition adds with sign to what can be a substantial elastic heat (*22*). The contribution of this elastic heat has not been universally appreciated but should become readily apparent when direct measurements are developed. The stress applied to BC materials is not necessarily perfectly isotropic in practice—for example, if the pressure-transmitting fluid is absent (*134*, *135*) or frozen—implying that mC effects could be decomposed into BC effects and eC effects. Prototype coolers are currently being developed by a U.K. spin-out company (*17*), and the key challenge is to exchange heat between source and sink via the cyclically pressurized BC material. Large pressures do not cause breakdown in most BC materials, and those that experience brittle fracture fortuitously present a large surface area for good heat exchange (*126*).

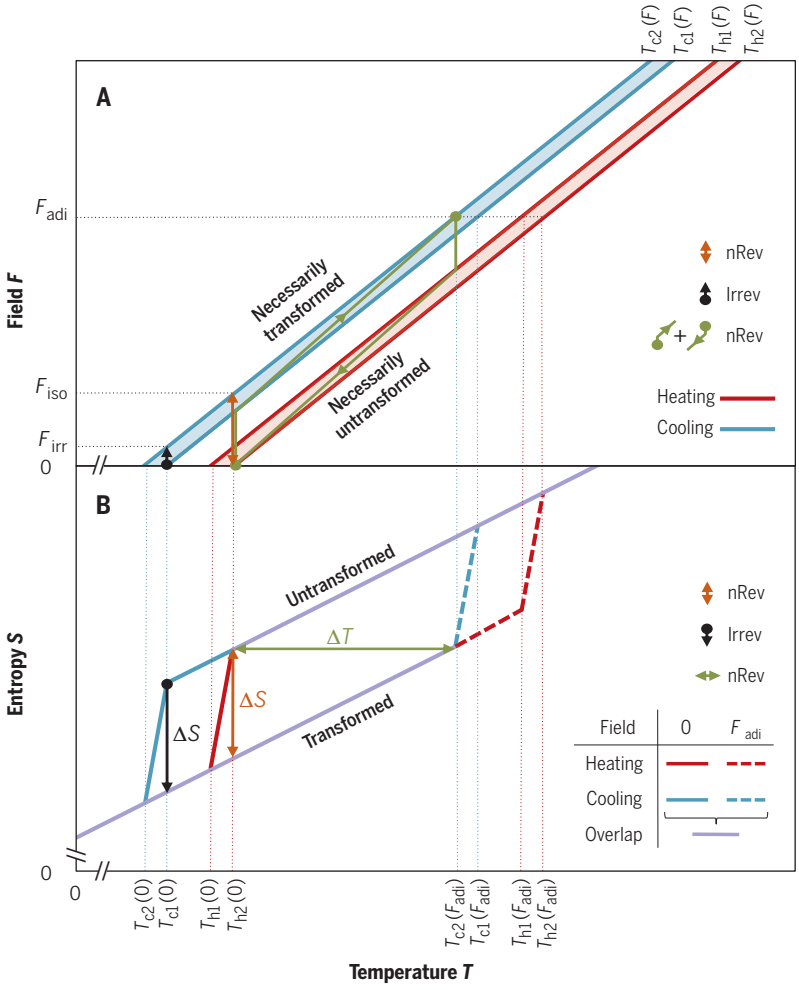


Fig. 5. Reversibility of caloric effects at a first-order transition far from the critical point. Hysteresis in both temperature and field are apparent when describing a conventional caloric material (*1*) via (**A**) a field-temperature (*F*-*T*) phase diagram and (**B**) the corresponding plot of entropy-temperature (*S*-*T*). (A) At constant *F*, heating causes the transition to start at $T_{h1}(F)$ and finish at $T_{h2}(F)$ (red lines), and cooling causes the transition to start at $T_{c1}(F)$ and finish at $T_{c2}(F)$ (blue lines) (the four lines could instead be labeled after isothermal field sweeps). The transformed fraction varies in the shaded regions, where we assume caloric effects to arise exclusively, such that the untransformed and transformed phases themselves show no caloric effects. (B) Hypothetical data obtained via heating (red lines) and cooling (blue lines) at zero field (solid lines) and F_{adi} (dashed lines) show partial overlap (purple lines). [(A) and (B)] Under isothermal conditions, the full transition can be driven in a nominally reversible (nRev) manner at minimum start temperature $T_{h2}(0)$ using F_{iso} (orange arrows). Also under isothermal conditions, the transition can be driven in an irreversible (Irrev) manner from the zero-field cooling branch either fully in $T_{c1}(0) \le T < T_{h2}(0)$ using a field as small as F_{irr} at $T_{c1}(0)$ (black arrows), or partially and to completion in $T_{c2}(0) < T < T_{c1}(0)$. Under adiabatic conditions, the full transition can be driven at minimum start temperature $T_{h2}(0)$ using F_{adi} (green arrows), and the process is nominally reversible. The full transition can (cannot) be driven in a nominally reversible manner at higher (any) start temperatures using fields that are larger (smaller) than F_{iso} and F_{adi}. If starting from the zero-field heating branch in $T_{h1}(0) < T < T_{h2}(0)$ (diagonal red solid line), then fields smaller than F_{iso} will partially or fully complete the transition in a nominally reversible manner.

Outlook

Given that heat pumps based on MC, EC, eC, and BC materials are now being developed toward commercial applications, we feel that it is helpful to provide a snapshot comparison by considering just one key advantage and one key disadvantage, while recognizing that MC prototypes are well ahead of the others in terms of performance. MC prototypes can easily achieve automatic energy recovery (*9*, *10*), but the field available from permanent magnets typically fails to drive MC effects very hard. EC working bodies are easy to drive with a voltage, but these working bodies tend to be relatively small in order to avoid electrical breakdown. Both eC and BC working bodies display large thermal changes, but eC working bodies are prone to breakdown, and BC working bodies require large driving pressures. Some advantage might therefore be gained from the growing body of work on multicaloric materials (*137*).

By definition, caloric effects in multicaloric materials can be driven by more than one type of driving field. For example, a change of stress can be used to complete a cooling cycle whose magnetic hysteresis unusually represents an advantage because it permits the transition to persist in a caloric material that has passed through a small region of magnetic field. This could permit both a reduction of permanent magnet volume and the concomitant possibility of field concentration (*138*). More simply, the combined use of magnetic and stress fields can broaden the range of operating temperatures by providing access to different transitions at

nearby temperatures (*139*). Elegant experimental excursions in parameter space threaten to increase the complexity of prototypes, unless the extra complexity instead resides wholly within the working body, as seen when using 0.16 T to magnetically drive 4 K changes in a shape memory alloy via magnetostrictive strain from juxtaposed Terfenol-D (*140*). This magnetically driven strain-mediated caloric effect in the shape memory alloy could conceivably be categorized as an MC effect or an eC effect, thus presenting a dichotomy with which to tax the taxonomist.

For now, most prototypes will likely continue to exploit monocaloric effects in monolithic materials and electrocaloric MLCs. Given the need for cyclical operation, it may be important to manage the flow of heat via thermal switches, thermal diodes, and heat pipes (*141*). Although it remains difficult to achieve large thermal conductivity contrast in solid devices with no moving parts, encouraging progress is reflected in the aforementioned electrostatically actuated films (*15*, *92*, *142*), the aforementioned mechanical heat switches (*89*, *90*), and a magnetically translated ferrofluid that transports heat (*143*). More generally, care should be taken to minimize not just inactive thermal mass but also any undue thermal resistance in prototypes, for example, within caloric materials and at interfaces such as corrosion-resistant coatings (*41*). The modeling of both materials and prototypes should play an important role in future developments, and we identify hysteresis as a key issue for both experimentalists and theorists to tackle, along with fatigue and breakdown, where relevant. It is also possible that caloric materials will find applications beyond cooling and heating, perhaps in energy harvesting based on plates of an MC material (*144*).

REFERENCES AND NOTES

1. X. Moya, S. Kar-Narayan, N. D. Mathur, *Nat. Mater.* **13**, 439–450 (2014).
2. G. V. Brown, *J. Appl. Phys.* **47**, 3673–3680 (1976).
3. C. R. H. Bahl *et al.*, *Int. J. Refrig.* **37**, 78–83 (2014).
4. T. Zhang, X.-S. Qian, H. Gu, Y. Hou, Q. M. Zhang, *Appl. Phys. Lett.* **110**, 243503 (2017).
5. M. O. McLinden, J. S. Brown, R. Brignoli, A. F. Kazakov, P. A. Domanski, *Nat. Commun.* **8**, 14476 (2017).
6. A. Kitanovski, U. Plaznik, U. Tomc, A. Poredoš, *Int. J. Refrig.* **57**, 288–298 (2015).
7. E. Söylemez, E. Alpman, A. Onat, *Int. J. Refrig.* **95**, 93–107 (2018).
8. B. Nair, thesis, University of Cambridge, Cambridge, UK (2020).
9. X. Moya, E. Defay, V. Heine, N. D. Mathur, *Nat. Phys.* **11**, 202–205 (2015).
10. C. Zimm *et al.*, *Int. J. Refrig.* **29**, 1302–1306 (2006).
11. M. Trček *et al.*, *Philos. Trans. R. Soc. London Ser. A* **374**, 20150301 (2016).
12. S. Qian *et al.*, *Sci. Technol. Built Environ.* **22**, 500–506 (2016).
13. F. Bruederlin, H. Ossmer, F. Wendler, S. Miyazaki, M. Kohl, *J. Phys. D Appl. Phys.* **50**, 424003 (2017).
14. E. Defay *et al.*, *Nat. Commun.* **9**, 1827 (2018).
15. Y. Meng *et al.*, *Nat. Energy* 10.1038/s41560-020-00715-3 (2020).
16. U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, "Energy Savings Potential and RD&D Opportunities for Commercial Building HVAC Systems" (2017); www.energy.gov/sites/prod/files/2017/12/f46/bto-DOE-Comm-HVAC-Report-12-21-17.pdf.
17. Henry Royce Institute, "Materials for the Energy Transition roadmap: Caloric Energy Conversion Materials" (2020); www.royce.ac.uk/content/uploads/2020/10/M4ET-Caloric-Energy-Conversion-Materials-roadmap.pdf.
18. B. Nair *et al.*, *Nature* **575**, 468–472 (2019).
19. S. Crossley, B. Nair, R. W. Whatmore, X. Moya, N. D. Mathur, *Phys. Rev. X* **9**, 041002 (2019).
20. C. R. H. Bahl, K. K. Nielsen, *J. Appl. Phys.* **105**, 013916 (2009).
21. E. Stern-Taulats *et al.*, arXiv:2010.15705 [cond-mat.mtrl-sci] (29 October 2020).
22. P. Lloveras *et al.*, *Nat. Commun.* **6**, 8801 (2015).
23. R. Z. Levitin, V. V. Snegirev, A. V. Kopylov, A. S. Lagutin, A. Gerber, *J. Magn. Magn. Mater.* **170**, 223–227 (1997).
24. S. Crossley, R. W. Whatmore, N. D. Mathur, X. Moya, arXiv:2010.15400 [cond-mat.mtrl-sci] (29 October 2020).
25. D. V. Christensen *et al.*, *J. Appl. Phys.* **108**, 063913 (2010).
26. J. Cui *et al.*, *Appl. Phys. Lett.* **101**, 073904 (2012).
27. M. G. Del Duca *et al.*, *J. Appl. Phys.* **128**, 104102 (2020).
28. S. Kar-Narayan *et al.*, *Appl. Phys. Lett.* **102**, 032903 (2013).
29. T. Usui *et al.*, *J. Phys. D Appl. Phys.* **50**, 424002 (2017).
30. M. Schmidt, A. Schütze, S. Seelecke, *Int. J. Refrig.* **54**, 88–97 (2015).
31. J. Tušek *et al.*, *Nat. Energy* **1**, 16134 (2016).
32. F. Bruederlin *et al.*, *Energy Technol.* **6**, 1588–1604 (2018).
33. F. Guillou, G. Porcari, H. Yibole, N. van Dijk, E. Brück, *Adv. Mater.* **26**, 2671–2675 (2014).
34. A. Chirkova *et al.*, *Acta Mater.* **131**, 31–38 (2017).
35. M. Fitta, R. Pełka, P. Konieczny, M. Balanda, *Crystals* **9**, 9 (2019).
36. V. Chaudhary, X. Chen, R. V. Ramanujan, *Prog. Mater. Sci.* **100**, 64–98 (2019).
37. B. Pulko *et al.*, *J. Magn. Magn. Mater.* **375**, 65–73 (2015).
38. I. A. Radulov, K. P. Skokov, D. Yu. Karpenkov, T. Gottschall, O. Gutfleisch, *J. Magn. Magn. Mater.* **396**, 228–236 (2015).
39. I. A. Radulov *et al.*, *Acta Mater.* **127**, 389–399 (2017).
40. H. Zhang *et al.*, *Scr. Mater.* **120**, 58–61 (2016).
41. J. Forchelet *et al.*, *Int. J. Refrig.* **37**, 307–313 (2014).
42. R. Yuan *et al.*, *J. Appl. Phys.* **127**, 154102 (2020).
43. T. Gottschall, K. P. Skokov, B. Frincu, O. Gutfleisch, *Appl. Phys. Lett.* **106**, 021901 (2015).
44. T. Gottschall, K. P. Skokov, R. Burriel, O. Gutfleisch, *Acta Mater.* **107**, 1–8 (2016).
45. H. Sepehri-Amin *et al.*, *Acta Mater.* **147**, 342–349 (2018).
46. J. Liu *et al.*, *Acta Mater.* **174**, 450–458 (2019).
47. M. Trassinelli *et al.*, *Appl. Phys. Lett.* **104**, 081906 (2014).
48. E. Lovell, A. M. Pereira, A. D. Caplin, J. Lyubina, L. F. Cohen, *Adv. Energy Mater.* **5**, 1401639 (2015).
49. M. G. Zavareh *et al.*, *Phys. Rev. Appl.* **8**, 014037 (2017).
50. C. Salazar Mejía *et al.*, *J. Appl. Phys.* **117**, 17E710 (2015).
51. N. Maraytta *et al.*, *J. Alloys Compd.* **805**, 1161–1167 (2019).
52. J. Döntgen *et al.*, *Appl. Phys. Lett.* **106**, 032408 (2015).
53. T. Gottschall *et al.*, *Phys. Rev. B* **99**, 134429 (2019).
54. J. Y. Law *et al.*, *Nat. Commun.* **9**, 2680 (2018).
55. M. E. Gruner *et al.*, *Phys. Rev. Lett.* **114**, 057202 (2015).
56. R. Bjørk, C. R. H. Bahl, K. K. Nielsen, *Int. J. Refrig.* **63**, 48–62 (2016).
57. S. Jacobs *et al.*, *Int. J. Refrig.* **37**, 84–91 (2014).
58. T. Lei, K. Engelbrecht, K. K. Nielsen, H. Neves Bez, C. R. H. Bahl, *J. Phys. D Appl. Phys.* **49**, 345001 (2016).
59. R. Teyber *et al.*, *Appl. Therm. Eng.* **106**, 405–414 (2016).
60. I. Niknia *et al.*, *J. Appl. Phys.* **123**, 204901 (2018).
61. T. Lei, K. Engelbrecht, K. K. Nielsen, C. T. Veje, *Appl. Therm. Eng.* **111**, 1232–1243 (2017).
62. J. Harada *et al.*, *Nat. Chem.* **8**, 946–952 (2016).
63. S. Crossley, W. Li, X. Moya, N. D. Mathur, *Philos. Trans. R. Soc. London Ser. A* **374**, 20150313 (2016).
64. W. Li *et al.*, *J. Mater. Chem. A* **8**, 16189–16194 (2020).
65. Y.-C. Zhao *et al.*, *ACS Omega* **4**, 14650–14654 (2019).
66. P. Vales-Castro *et al.*, arXiv:2009.02184 [cond-mat.mtrl-sci] (4 September 2020).
67. J. Li *et al.*, *Acta Mater.* **115**, 58–67 (2016).
68. X.-S. Qian *et al.*, *Adv. Funct. Mater.* **24**, 1300–1305 (2014).
69. J. Koruza, B. Rožič, G. Cordoyiannis, B. Malič, Z. Kutnjak, *Appl. Phys. Lett.* **106**, 202905 (2015).
70. A.-K. Axelsson, F. Le Goupil, M. Valant, N. McN. Alford, *Acta Mater.* **124**, 120–126 (2017).
71. C. A.-P. de Araujo, J. D. Cuchiaro, L. D. McMillan, M. C. Scott, J. F. Scott, *Nature* **374**, 627–629 (1995).
72. H.-J. Ye *et al.*, *Appl. Phys. Lett.* **105**, 152908 (2014).
73. J. Li *et al.*, *ACS Appl. Mater. Interfaces* **11**, 23346–23352 (2019).
74. B. L. Liu *et al.*, *RSC Advances* **4**, 24533–24537 (2014).
75. S. Crossley *et al.*, *Appl. Phys. Lett.* **108**, 032902 (2016).
76. S. Pandya *et al.*, *Phys. Rev. Appl.* **7**, 034025 (2017).
77. S. G. Lu *et al.*, *Appl. Phys. Lett.* **97**, 202901 (2010).
78. G. Zhang *et al.*, *Adv. Mater.* **27**, 1450–1454 (2015).
79. Q. Li *et al.*, *Adv. Mater.* **27**, 2236–2241 (2015).
80. G. Zhang *et al.*, *Adv. Mater.* **31**, e1806642 (2019).
81. L. Fulanović *et al.*, *J. Eur. Ceram. Soc.* **37**, 599–603 (2017).
82. M. Almanza, J. Chavanne, Y. Civet, Y. Perriard, *Smart Mater. Struct.* **29**, 045044 (2020).
83. B. Kwon *et al.*, *Appl. Phys. Lett.* **104**, 213902 (2014).
84. L. Fulanović *et al.*, *J. Eur. Ceram. Soc.* **37**, 5105–5108 (2017).
85. A. Bradeško *et al.*, *Acta Mater.* **169**, 275–283 (2019).
86. F. Weyland *et al.*, *J. Eur. Ceram. Soc.* **38**, 551–556 (2018).
87. D. Sette *et al.*, *APL Mater.* **4**, 091101 (2016).
88. A. Torelló *et al.*, *Science* **370**, 125–129 (2020).
89. Y. D. Wang *et al.*, *Appl. Phys. Lett.* **107**, 134103 (2015).
90. Y. Wang *et al.*, *Science* **370**, 129–133 (2020).
91. Y. V. Sinyavskii, *Chem. Petrol. Eng.* **31**, 295–306 (1995).
92. R. Ma *et al.*, *Science* **357**, 1130–1134 (2017).
93. Z. Xie, G. Sebald, D. Guyomar, *Phys. Lett. A* **381**, 2112–2116 (2017).
94. Y. Yoshida, K. Yuse, D. Guyomar, J.-F. Capsal, G. Sebald, *Appl. Phys. Lett.* **108**, 242904 (2016).
95. S. S. Hong *et al.*, *Science* **368**, 71–76 (2020).
96. Y. Wu, E. Ertekin, H. Sehitoglu, *Acta Mater.* **135**, 158–176 (2017).
97. F. Xiao, M. Jin, J. Liu, X. Jin, *Acta Mater.* **96**, 292–300 (2015).
98. G. J. Pataky, E. Ertekin, H. Sehitoglu, *Acta Mater.* **96**, 420–427 (2015).
99. F. Xiao, T. Fukuda, T. Kakeshita, X. Jin, *Acta Mater.* **87**, 8–14 (2015).
100. K. Otsuka, X. Ren, *Prog. Mater. Sci.* **50**, 511–678 (2005).
101. X. Liang *et al.*, *Scr. Mater.* **134**, 42–46 (2017).
102. F. Xiao, T. Fukuda, T. Kakeshita, *Scr. Mater.* **124**, 133–137 (2016).
103. D. Cong *et al.*, *Phys. Rev. Lett.* **122**, 255703 (2019).
104. J. A. Shaw, S. Kyriakides, *J. Mech. Phys. Solids* **43**, 1243–1281 (1995).
105. E. A. Pieczyska, S. P. Gadaj, W. K. Nowacki, H. Tobushi, *Exp. Mech.* **46**, 531–542 (2006).
106. H. Hou *et al.*, *J. Phys. D Appl. Phys.* **50**, 404001 (2017).
107. Z. Xie, G. Sebald, D. Guyomar, *Appl. Phys. Lett.* **107**, 081905 (2015).
108. G. Sebald, Z. Xie, D. Guyomar, *Philos. Trans. R. Soc. London Ser. A* **374**, 20150302 (2016).
109. F. Xiao *et al.*, *Acta Mater.* **118**, 88–94 (2016).
110. H. Hou *et al.*, *Science* **366**, 1116–1121 (2019).
111. Y. Xu, B. Lu, W. Sun, A. Yan, J. Liu, *Appl. Phys. Lett.* **106**, 201903 (2015).
112. Z. Yang *et al.*, *Mater. Res. Lett.* **7**, 137–144 (2019).
113. L. Wei, X. Zhang, W. Gan, C. Ding, L. Geng, *Scr. Mater.* **168**, 28–32 (2019).
114. H. Chen *et al.*, *Acta Mater.* **158**, 330–339 (2018).
115. C. Chluba, H. Ossmer, C. Zamponi, M. Kohl, E. Quandt, *Shape Memory and Superelasticity* **2**, 95–103 (2016).
116. Z. Yang, D. Y. Cong, X. M. Sun, Z. H. Nie, Y. D. Wang, *Acta Mater.* **127**, 33–42 (2017).
117. J. Tušek *et al.*, *Acta Mater.* **150**, 295–307 (2018).
118. L. Wei, X. Zhang, J. Liu, L. Geng, *AIP Adv.* **8**, 055312 (2018).
119. K. Engelbrecht *et al.*, *APL Mater.* **4**, 064110 (2016).
120. S. Qian *et al.*, *Int. J. Refrig.* **64**, 1–19 (2016).
121. R. Snodgrass, D. Erickson, *Sci. Rep.* **9**, 18532 (2019).
122. R. Wang *et al.*, *Science* **366**, 216–221 (2019).
123. D. Matsunami, A. Fujita, K. Takenaka, M. Kano, *Nat. Mater.* **14**, 73–78 (2015).
124. T. Samanta *et al.*, *Appl. Phys. Lett.* **112**, 021907 (2018).
125. D. Boldrin *et al.*, *Phys. Rev. X* **8**, 041035 (2018).
126. A. Aznar *et al.*, *Adv. Mater.* **31**, e1903577 (2019).
127. E. Stern-Taulats *et al.*, *APL Mater.* **4**, 091102 (2016).
128. A. Aznar *et al.*, *Nat. Commun.* **8**, 1851 (2017).
129. M. V. Gorev, E. A. Mikhaleva, I. N. Flerov, E. V. Bogdanov, *J. Alloys Compd.* **806**, 1047–1051 (2019).
130. I. N. Flerov, M. V. Gorev, A. V. Kartashev, E. I. Pogorel'tsev, N. M. Laptash, *J. Mater. Sci.* **54**, 14287–14295 (2019).
131. J. M. Bermúdez-García *et al.*, *Nat. Commun.* **8**, 15715 (2017).
132. J. M. Bermúdez-García *et al.*, *J. Mater. Chem. C* **6**, 9867–9874 (2018).
133. S. P. Vallone *et al.*, *Adv. Mater.* **31**, e1807334 (2019).
134. E. O. Usuda, N. M. Bom, A. M. G. Carvalho, *Eur. Polym. J.* **92**, 287–293 (2017).
135. A. M. G. Carvalho, W. Imamura, E. O. Usuda, N. M. Bom, *Eur. Polym. J.* **99**, 212–221 (2018).
136. P. Lloveras *et al.*, *Nat. Commun.* **10**, 1803 (2019).
137. E. Stern-Taulats *et al.*, *MRS Bull.* **43**, 295–299 (2018).
138. T. Gottschall *et al.*, *Nat. Mater.* **17**, 929–934 (2018).
139. Y. Hu *et al.*, *APL Mater.* **5**, 046103 (2017).
140. H. Hou, P. Finkel, M. Staruch, J. Cui, I. Takeuchi, *Nat. Commun.* **9**, 4075 (2018).
141. K. Klinar, A. Kitanovski, *Renew. Sustain. Energy Rev.* **118**, 109571 (2020).
142. M. Almanza, L. Depreux, F. Parrain, M. LoBue, *Appl. Phys. Lett.* **112**, 083901 (2018).
143. J. B. Puga *et al.*, *Nano Energy* **31**, 278–285 (2017).
144. A. Waske *et al.*, *Nat. Energy* **4**, 68–74 (2019).

ACKNOWLEDGMENTS

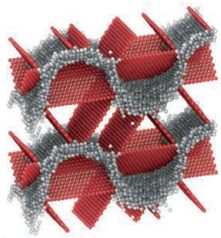
We thank B. Nair and E. Defay for helpful discussions and S. Crossley for making the link between EC measurements and gambling. **Funding:** X.M. received support from ERC Starting Grant 680032 and the Royal Society. **Competing interests:** X.M. is director of research at Barocal Ltd.

10.1126/science.abb0973

RESEARCH

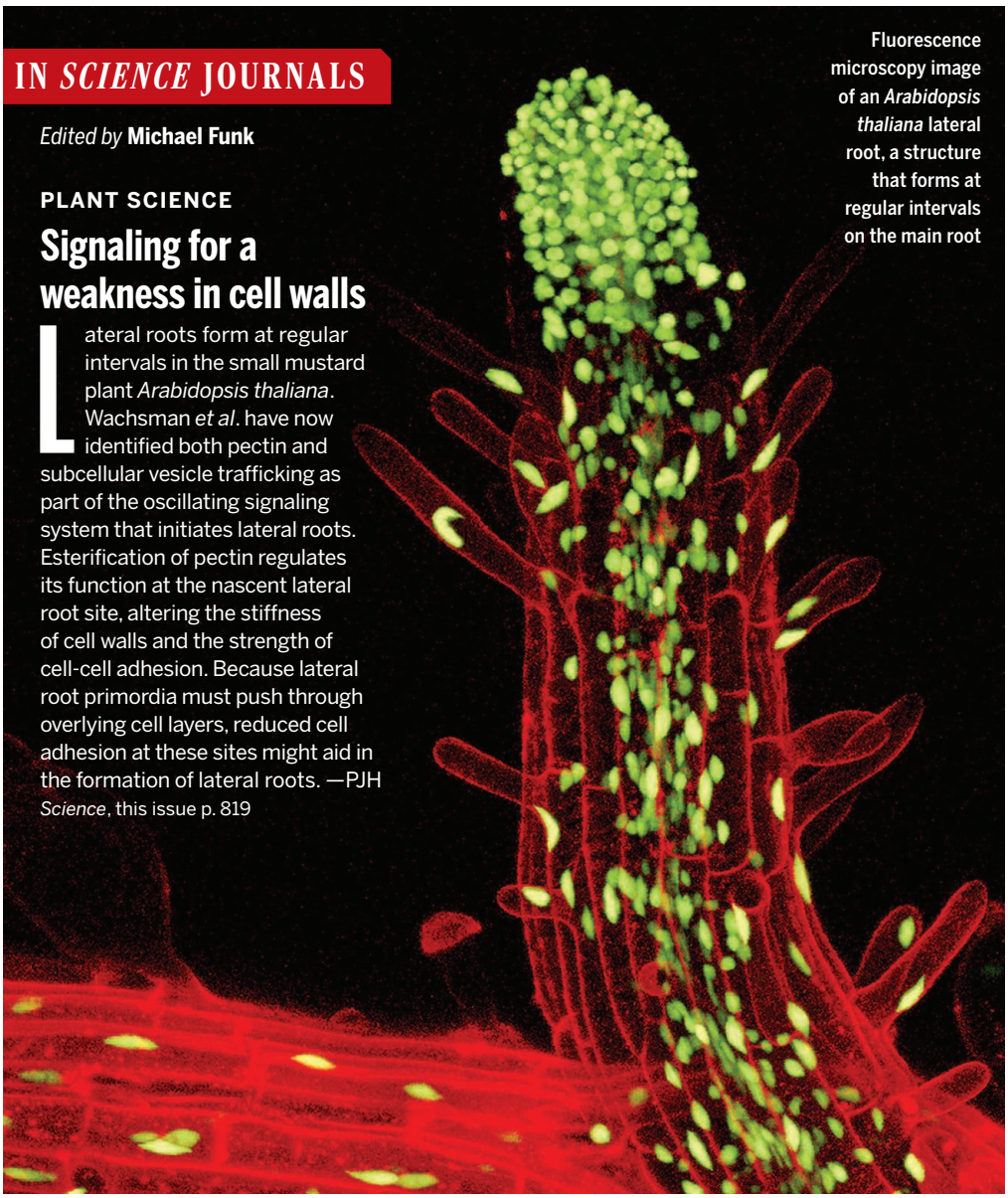
Temperature-stable nanograined copper

Li et al., *p. 831*



IN *SCIENCE* JOURNALS

Edited by **Michael Funk**



Fluorescence microscopy image of an *Arabidopsis thaliana* lateral root, a structure that forms at regular intervals on the main root

PLANT SCIENCE

Signaling for a weakness in cell walls

Lateral roots form at regular intervals in the small mustard plant *Arabidopsis thaliana*. Wachsman *et al*. have now identified both pectin and subcellular vesicle trafficking as part of the oscillating signaling system that initiates lateral roots. Esterification of pectin regulates its function at the nascent lateral root site, altering the stiffness of cell walls and the strength of cell-cell adhesion. Because lateral root primordia must push through overlying cell layers, reduced cell adhesion at these sites might aid in the formation of lateral roots. —PJH

Science, this issue p. 819

STRESS RESPONSES

Cellular adaptation during metabolic stress

Cells respond to environmental stress by down-regulating general protein synthesis and inducing selective expression of proteins required for survival. However, the mechanisms controlling this selective messenger RNA (mRNA) translation response remain poorly understood. Lamper *et al*. report that the noncanonical 5′ cap-binding protein subunit eIF3d is activated upon metabolic stress in mammalian cells to reprogram cellular mRNA translation. eIF3d is activated by a switch in phosphorylation status at sites near the cap-binding pocket and enables cells to express the proteins required for the regulation of metabolism and survival during stresses, including glucose starvation. This work reveals how eIF3d-dependent, noncanonical cap-dependent translation controls the cellular adaptation to stress. —SMH

Science, this issue p. 853

MARTIAN ATMOSPHERE

Dust storms cause Mars to lose water

Mars was once a wet planet, but it has lost most of its water through reactions that produce hydrogen, which escapes from the upper atmosphere into space. Stone *et al*. used data from the Mars Atmosphere and Volatile Evolution spacecraft to study how water is transported to the upper atmosphere and converted to hydrogen. They found that water can reach higher altitudes than previously thought, especially during global or regional dust storms. Photochemical modeling shows that this process dominates the current loss of water from Mars and influenced the evolution of its climate. —KTS

Science, this issue p. 824

SENSORS

Colorful changes

Distributed fiber-optic sensors have been used for monitoring mechanical deformations in stiff infrastructures such as bridges, roads, and buildings, but they either are limited to measuring one variable or require complex optics to measure multiple properties. Bai *et al*. now demonstrate dual-core elastomeric optical fibers, one of which contains patterned dye regions. The waveguides are fabricated by molding out of commercially available elastomers and integrate a clear core and an adjacent core doped with up to three macroscale dye regions. Changes in optical paths in the two cores detect deformation and map it onto a color space. By monitoring changes in the color and intensity in both

IMAGE: DR. JOHN RUNIONS/SCIENCE SOURCE

Published by AAAS

elastomer-based fibers, the researchers could distinguish bending, stretching, and localized pressing with a spatial resolution down to ~1 centimeter. —MSL

Science, this issue p. 848

OPTOMECHANICS

Getting phonons to hang around

The ideal platform for quantum-computing and quantum-sensing applications is likely to be a hybrid system that combines the best features of different components. Superconducting circuits are relatively advanced, and finding components that can control and manipulate the microwaves will be essential. MacCabe *et al*. explored the use of high-quality microresonators in which the acoustic environment could be engineered such that the phonon lifetime could be extended to more than 1 second. Operating at microwave frequencies of 5 gigahertz, these quantum acoustic-dynamic devices could be coupled with superconducting circuits. —ISO

Science, this issue p. 840

VNEUROSCIENCE

Higher-order thalamus input to the cortex

Sensory information can only be used meaningfully in the brain when integrated with and compared with internally generated top-down signals. However, we know little about the brainwide afferents that convey such top-down signals, their information content, and learning-related plasticity. Pardi *et al*. identified the higher-order thalamus as a major source of top-down input to mouse auditory cortex and investigated a circuit in cortical layer 1 that facilitates plastic changes and flexible responses. These results demonstrate how top-down feedback information can reach cortical areas through a noncortical structure that has received little attention despite its widespread connections with the cortex. —PRS

Science, this issue p. 844

STROKE

Measuring brain damage using the blood

The outcome of a stroke varies greatly between patients, from temporary mild symptoms to permanent disability and death. The clinical scale for measuring disease severity shows poor correlation with brain tissue damage. Identification of better markers of tissue damage could improve the ability to predict outcome and promote the development of better therapies. Gendron *et al*. found that the presence of the axonal cytoskeletal protein neurofilament light (NFL) in the blood was increased in multiple cohorts of patients with stroke compared with controls and was correlated with brain tissue damage. Additionally, blood concentration of NFL was correlated with functional outcome and mortality. These results suggest that blood NFL might be used as a prognostic marker after stroke. —MM

Sci. Transl. Med. **12**, eaay1913 (2020).

ANTIVIRAL IMMUNITY

Role reversal for aged T cells

Age-related changes to the adaptive immune system are associated with impaired host response and increased morbidity and mortality after respiratory virus infection. Goplen *et al*. used a murine model of influenza virus infection to compare the function of lung tissue–resident memory T (TRM) cells in young and aged mice. Although TRM cells typically promote robust antiviral immunity, aged mice displayed increased accumulation of dysfunctional, influenza-specific TRM cells. These cells facilitated pathological lung fibrosis and provided less protection against heterologous infection compared with TRM cells from young mice. This study provides insight into how age and prior infection may affect T cell–mediated immunity to respiratory viruses. —CO

Sci. Immunol. **5**, eabc4557 (2020).

IN OTHER JOURNALS

Edited by **Caroline Ash** *and* **Jesse Smith**

IMAGE: NOAA VIA GETTY IMAGES

BIOMATERIALS

Cell spreading affects energy consumption

The stiffness of a substrate is known to affect how cells spread on its surface, and for stem cells, stiffness also can affect their proliferation pathway. Xie *et al*. attempted to quantify another aspect of cell behavior, the influence of matrix stiffness on the availability of energy expenditure. Stiffer substrates cause the formation of focal adhesions and reorganization of the actin cytoskeleton. This leads to an intracellular drop in adenosine triphosphate (ATP) levels and the subsequent activation of adenosine monophosphate–activated protein kinase (AMPK) and glucose uptake to produce ATP and to support cell tension and adhesion. AMPK activation altered the mitochondrial morphology, causing it to be more fragmented. When the authors tested AMPKα-null cells, they observed limited spreading, a less well-developed actin cytoskeleton, and fewer focal adhesions on stiff substrates relative to the wild type. —MSL

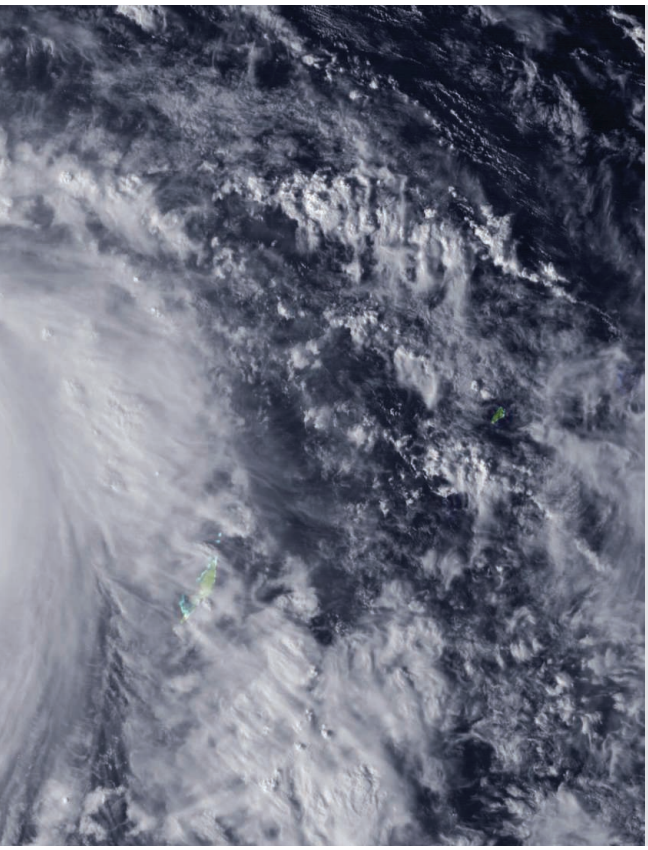
Biomaterials **267**, 120494 (2020).

STRUCTURAL BIOLOGY

Mapping a dynamic structural ensemble

Many biological complexes are conformationally dynamic, which makes it difficult to determine the structures required to understand the mechanism. Marx *et al*. used an integrative approach to study the dynamic process of outer membrane protein (OMP) biogenesis that occurs in the periplasm of Gram-negative bacteria. Using a combination of photo–cross-linking, mass spectroscopy, solution scattering, and molecular modeling, they mapped interactions of a key chaperone, SurA, with unfolded OMPs (uOMPs) and determined an ensemble of models of the SurA:uOMP complex. The data show that a groove in SurA can bind several regions in uOMPs. Binding results in an expansion of the rest of the uOMP. Structural models indicate how

Published by AAAS



TROPICAL CYCLONES

Birth of a storm

How do tropical cyclones form from smaller tropical disturbances? Only a fraction of tropical disturbances evolve into the highly energetic monsters called cyclones or hurricanes, so being able to predict which ones will make the transition could help improve forecasts. Ruppert *et al*. show that infrared radiation trapped by the deep clouds found in these precursor storms causes local heating that is key for their development into full-blown tropical cyclones. Incorporating these insights about the role of cloud-radiation feedback into numerical models could lead to better representations of tropical cyclone genesis and intensification. —HJS

Proc. Natl. Acad. Sci. U.S.A. 10.1073/pnas.2013584117 (2020).

Satellite image of Super Typhoon Haiyan as it approaches the Philippines on 7 November 2013.

SurA prevents OMP aggregation and how it may deliver uOMP to the β-barrel assembly machinery complex for insertion into the outer membrane. —VV

Proc. Natl. Acad. Sci. U.S.A. 10.1073/pnas.2008175117 (2020).

PLANT ECOLOGY

Heating up an invasion

Populations of invasive plant species are likely to grow more rapidly in a warming climate. To test this concept, Keller and Shea experimentally studied the growth and reproduction of the globally invasive thistle *Carduus nutans* in warmed plots. Both the temperature and the onset of the growing season were then manipulated. Their results and modeling show that plants reached reproductive maturity earlier and grew larger, which enabled the production of more seeds. Overall, the thistle's population growth rate was projected to increase by 15%, posing a further threat to agricultural production and food security in regions where it has already become invasive. —AMS

Ecology 10.1002/ECY.3219 (2020).

PHOTO: ADVENTURE_PHOTO/ISTOCKPHOTO

The population growth rate of the invasive thistle *Carduus nutans* is projected to increase as the climate heats up.

CANCER

Precision therapy for leukemia

Acute myeloid leukemia is an aggressive cancer that is typically treated immediately. Such rapid treatment responses do not allow time for detailed genetic analysis. To test the safety of delaying treatment and the utility of genetic analysis in this patient population, Burd *et al*. performed a clinical trial in 395 older adults with acute myeloid leukemia. A 7-day delay in treatment was sufficient for genetic testing and safe for most subgroups of patients. In addition, the individualized treatment guided by genetic testing greatly improved patient survival, supporting the potential utility of this approach for clinical care. —YN

Nat. Med. 10.1038/s41591-020-1089-8 (2020).

EXERCISE

Bioenergetic sensor for exercise

Exercising provides multiple benefits to the body by increasing strength, reducing disease, managing weight, and improving brain health. Using a metabolomics approach, Reddy *et al*. show that succinate is secreted in mice and humans during exercise by a pH-gated mechanism involving monocarboxylate transporter 1 (MCT1). The succinate that is released during exercise signals through the succinate receptor 1 (SUCNR1) to regulate the paracrine response and affect innervation and extracellular matrix remodeling in muscle. Further, SUCNR1 signaling antagonizes muscle inflammation. This work shows that paracrine regulation by the succinate-SUCNR1 pathway in muscle is important for muscle strength and innervation, as well as for restoring insulin sensitivity. —BAP

Cell **183**, 62 (2020).

FIRE RESEARCH

Following the firebrands

Wildfires spread in a number of ways, and understanding how they do is particularly important for wildland-urban interface fires. Suzuki and Manzello used a specialized experimental protocol to simulate fire spread through firebrands, the various-sized embers ejected during combustion. The authors found that changing wind speeds had an impact on new ignition from firebrands. Understanding the interplay between direct flame contact, thermal radiation, and firebrand character is important for better predicting wildfire spread. —BG

Fire Technol. 10.1007/s10694-020-01018-5 (2020).

Published by AAAS

ALSO IN *SCIENCE* JOURNALS

Edited by **Michael Funk**

CORONAVIRUS

Estimating vaccine efficacy

Numerous coronavirus disease 2019 (COVID-19) vaccines are under development in clinical trials, but what do we need to know to ensure that they are used correctly? There are two vaccination strategies: direct protection of vulnerable populations and vaccination of the general population to indirectly protect those at highest risk of severe disease. In a Perspective, Lipsitch and Dean outline the important features of vaccine responses, particularly efficacy in high-risk groups and reduction of infectiousness. These end points are not routinely examined in phase 3 trials, so it is important to develop trials and subsequent studies to assess how vaccines perform so that they can be used optimally. —GKA

Science, this issue p. 763

HUMAN GENOMICS

The genomics of human development

Understanding the trajectory of a developing human requires an understanding of how genes are regulated and expressed. Two papers now present a pooled approach using three levels of combinatorial indexing to examine the single-cell gene expression and chromatin landscapes from 15 organs in fetal samples. Cao *et al*. focus on measurements of RNA in broadly distributed cell types and provide insights into organ specificity. Domcke *et al*. examined the chromatin accessibility of cells from these organs and identify the regulatory elements that regulate gene expression. Together, these analyses generate comprehensive atlases of early human development. —LMZ

Science, this issue p. 808, p. 809

PLANT SCIENCE

Roots primed for better phosphate uptake

Phosphate is a key resource for plants, and remediating phosphate deficiency drives considerable fertilizer use. In low-phosphate conditions, roots make more root hairs, which makes them better able to take up what little phosphate can be found. Wendrich *et al*. performed single-cell transcriptomics on the developing *Arabidopsis* root and queried the resulting gene-expression atlas for responses related to vascular development. The authors found that signals regulating root hair development began in the inner vasculature of the root with transcription factors that drove the production of the hormone cytokinin. Response cascades identified through the transcriptome database pointed to genes in epidermal cells that regulate root hair development. —PJH

Science, this issue p. 810

CORONAVIRUS

Imperfect future immunity

Humans are infected by several seasonal and cross-reacting coronaviruses. None provokes fully protective immunity, and repeat infections are the norm. Vaccines tend to be less efficient than natural infections at provoking immunity, and there are risks of adverse cross-reactions. Saad-Roy *et al*. used a series of simple models for a variety of immune scenarios to envisage immunological futures for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with and without vaccines. The model outcomes show that our imperfect knowledge about the imperfect coronavirus immune landscape can give rise to diverging scenarios ranging from recurring severe epidemics to elimination. It is critical that we accurately characterize immune responses to SARS-CoV-2 for translation into managing disease control. —CA

Science, this issue p. 811

COSMOCHEMISTRY

Timing Solar System formation

The oldest solids that formed in the Solar System are calcium-aluminium–rich inclusions (CAIs), small metallic droplets that were later incorporated into meteorites. The ages of CAIs are conventionally taken as the age of the Solar System, but which exact moment in star formation they correspond to has been unclear. Brennecka *et al*. measured molybdenum isotope ratios in CAIs and found a wide range of origins in both the inner and outer Solar System. They propose that CAIs formed from heterogeneous material accreting from the presolar nebula and that the ages of CAIs coincide with the Sun's transition from a protostar to a pre–main sequence star. —KTS

Science, this issue p. 837

CORONAVIRUS

Another host factor for SARS-CoV-2

Virus-host interactions determine cellular entry and spreading in tissues. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the earlier SARS-CoV use angiotensin-converting enzyme 2 (ACE2) as a receptor; however, their tissue tropism differs, raising the possibility that additional host factors are involved. The spike protein of SARS-CoV-2 contains a cleavage site for the protease furin that is absent from SARS-CoV (see the Perspective by Kielian). Cantuti-Castelvetri *et al*. now show that neuropilin-1 (NRP1), which is known to bind furin-cleaved substrates, potentiates SARS-CoV-2 infectivity. NRP1 is abundantly expressed in the respiratory and olfactory epithelium, with highest expression in endothelial and epithelial cells. Daly *et al*. found that the furin-cleaved S1 fragment of the spike protein binds directly to cell surface NRP1 and blocking this interaction with a small-molecule inhibitor or monoclonal antibodies reduced viral infection in cell culture. Understanding the role of NRP1 in SARS-CoV-2 infection may suggest potential targets for future antiviral therapeutics. —SMH

Science, this issue p. 856, p. 861; see also p. 765

PREBIOTIC CHEMISTRY

Cysteine as peptide precursor and catalyst

Among amino acids, cysteine is highly reactive as a nucleophile, metal ligand, and participant in redox and radical reactions. These properties make cysteine attractive as a component of prebiotic chemistry, but traditional Strecker synthesis of α-aminonitriles, which can serve as peptide precursors, cannot produce free cysteine. Foden *et al*. found that a simple acylation of the free amine prevented degradation of cysteine nitrile and enabled synthesis of this cysteine precursor from acetyl dehydroalanine nitrile and a sulfide donor (see the Perspective by Muchowska and Moran). When combined with other proteinogenic α-aminonitriles, acetylcysteine or derivative thiols catalyzed efficient peptide ligation in water. These results highlight how prebiotic synthesis of precursors can also generate function by creating a catalyst for polymerization. —MAF

Science, this issue p. 865; see also p. 767

Published by AAAS

MOLECULAR GENETICS

Circular RNAs protect male fertility

Eukaryotes make thousands of circular RNAs (circRNAs) by noncanonical RNA splicing. The functions of most are mysterious, but many sequester microRNAs or RNA-binding proteins. Gao *et al*. examined the functions of circRNAs from the conserved reproductive gene *BOULE* (circBoule RNAs). Loss of these RNAs reduced male fertility in fruit flies and mice under heat stress, with reduced sperm levels and altered sperm morphology. During spermatogenesis, circBoule RNAs interact with heat shock proteins and control their levels by promoting their ubiquitination. circBoule RNA interaction with heat shock proteins is conserved in human sperm, and low circBoule RNA levels correlate with low sperm motility. These findings reveal conserved molecular and physiological functions of circBoule RNAs over some 600 million years of evolution. —DD

Sci. Adv. 10.1126/sciadv.abb7426 (2020).

CANCER

Converging on DKK1 to drive metastasis

Hepatocellular carcinoma (HCC) is a common form of liver cancer. The Wnt signaling protein Dickkopf-1 (DKK1) and epidermal growth factor receptor (EGFR) are both associated with metastatic progression and poor prognosis in HCC patients. Niu *et al*. found that these molecular markers are linked. The activation of EGFR in HCC cells induced *DKK1* expression through parallel pathways that promoted nuclear translocation of the kinase PKM2 and activation of the acetyltransferase p300. These pathways converged on modifying histone H3 at the *DKK1* promoter to activate *DKK1* transcription. —LKF

Sci. Signal. **13**, eabb5727 (2020).

METALLURGY

Locking in nanoscale strength

Metals with nanometer-sized crystal grains are super strong, but they do not generally retain their structure at higher temperatures. This property undermines their high strength and makes their use in applications challenging. Li *et al*. found a minimum-interface structure in copper with 10-nanometer-sized grains that, when combined with a nanograin crystallographic twinning network, retains high strength to temperatures just below the melting point. This discovery suggests a different path forward for stabilizing nanograined metals. —BG

Science, this issue p. 831

Published by AAAS

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

A human cell atlas of fetal gene expression

Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Paul D. Kingsley, Mei Deng, Riza M. Daza, Michael A. Zager, Kimberly A. Aldinger, Ronnie Blecher-Gonen, Fan Zhang, Malte Spielmann, James Palis, Dan Doherty, Frank J. Steemers, Ian A. Glass, Cole Trapnell*, Jay Shendure*

INTRODUCTION: A reference atlas of human cell types is a major goal for the field. Here, we set out to generate single-cell atlases of both gene expression (this study) and chromatin accessibility (Domcke *et al.*, this issue) using diverse human tissues obtained during midgestation.

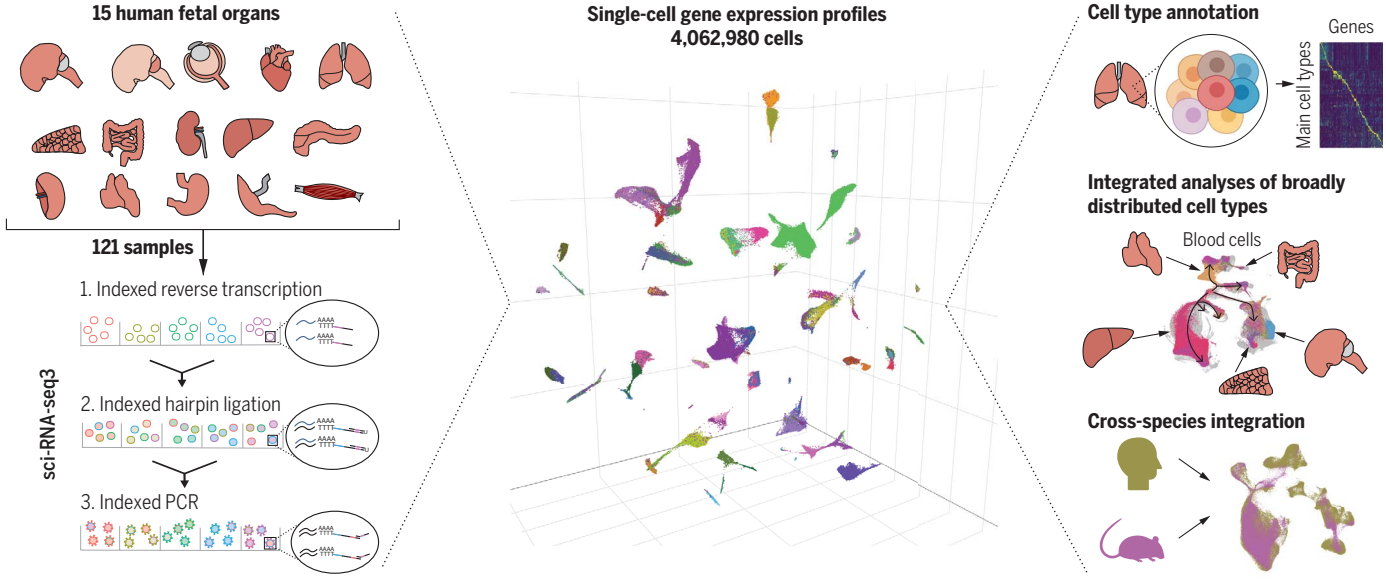
RATIONALE: Contemporary knowledge of the molecular basis of in vivo human development mostly derives from a combination of human genetics, in vivo investigations of model organisms, and in vitro studies of differentiating human cell lines, rather than through direct investigations of developing human tissues. Several challenges have historically limited the study of developing human tissues at the molecular level, including limited access, tissue degradation, and cell type heterogeneity. For this and the companion study (Domcke *et al.*, this issue), we were able to overcome these challenges.

RESULTS: We applied three-level single-cell combinatorial indexing for gene expression (sci-RNA-seq3) to 121 human fetal samples ranging from 72 to 129 days in estimated postconceptual age and representing 15 organs, altogether profiling 4 million single cells. We developed and applied a framework for quantifying cell type specificity, identifying 657 cell subtypes, which we preliminarily annotated based on cross-matching to mouse cell atlases. We identified and validated potentially circulating trophoblast-like and hepatoblast-like cells in unexpected tissues. Profiling gene expression in diverse tissues facilitated the cross-tissue analyses of broadly distributed cell types, including blood, endothelial, and epithelial cells. For blood cells, this yielded a multiorgan map of cell state trajectories from hematopoietic stem cells to all major sublineages. Multiple lines of evidence support the adrenal gland as a normal, albeit minor, site of erythropoiesis during fetal development. It was notably straightforward to integrate these human fetal data with a mouse embryonic cell atlas, despite differences in species and developmental stage. For some systems, this essentially permitted us to bridge gene expression dynamics from the embryonic to the fetal stages of mammalian development.

CONCLUSION: The single-cell data resource presented here is notable for its scale, its focus on human fetal development, the breadth of tissues analyzed, and the parallel generation of gene expression (this study) and chromatin accessibility data (Domcke *et al.*, this issue). We furthermore consolidate the technical framework for individual laboratories to generate and analyze gene expression and chromatin accessibility data from millions of single cells. Looking forward, we envision that the somewhat narrow window of midgestational human development studied here will be complemented by additional atlases of earlier and later time points, as well as similarly comprehensive profiling and integration of data from model organisms. The continued development and application of methods for ascertaining gene expression and chromatin accessibility—in concert with spatial, epigenetic, proteomic, lineage history, and other information—will be necessary to obtain a comprehensive view of the temporal unfolding of human cell type diversity that begins at the single-cell zygote. An interactive website facilitates the exploration of these freely available data by tissue, cell type, or gene (descartes.brotmanbaty.org). ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: coletrap@uw.edu (C.T.); shendure@uw.edu (J.S.)
Cite this article as J. Cao *et al.*, *Science* **370**, eaba7721 (2020). DOI: 10.1126/science.aba7721

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.aba7721



A human cell atlas of fetal gene expression enables the exploration of in vivo gene expression across diverse cell types. We used a three-level combinatorial indexing assay (sci-RNA-seq3) to profile gene expression in ~4,000,000 single cells from 15 fetal organs. This rich resource enables, for example, the identification and annotation of cell types, cross-tissue integration of broadly distributed cell types (e.g., blood, endothelial, and epithelial), and interspecies integration of mouse embryonic and human fetal cell atlases. PCR, polymerase chain reaction.

RESEARCH ARTICLE

HUMAN GENOMICS

A human cell atlas of fetal gene expression

Junyue Cao[1]*, Diana R. O'Day[2], Hannah A. Pliner[3], Paul D. Kingsley[4], Mei Deng[2], Riza M. Daza[1], Michael A. Zager[3,5], Kimberly A. Aldinger[2,6], Ronnie Blecher-Gonen[1], Fan Zhang[7], Malte Spielmann[8,9], James Palis[4], Dan Doherty[2,3,6], Frank J. Steemers[7], Ian A. Glass[2,3,6], Cole Trapnell[1,3,10]†, Jay Shendure[1,3,10,11]†

The gene expression program underlying the specification of human cell types is of fundamental interest. We generated human cell atlases of gene expression and chromatin accessibility in fetal tissues. For gene expression, we applied three-level combinatorial indexing to >110 samples representing 15 organs, ultimately profiling ~4 million single cells. We leveraged the literature and other atlases to identify and annotate hundreds of cell types and subtypes, both within and across tissues. Our analyses focused on organ-specific specializations of broadly distributed cell types (such as blood, endothelial, and epithelial), sites of fetal erythropoiesis (which notably included the adrenal gland), and integration with mouse developmental atlases (such as conserved specification of blood cells). These data represent a rich resource for the exploration of in vivo human gene expression in diverse tissues and cell types.

To date, most investigations of human development have been anatomical or histological in nature (*1*–*3*). However, it is clear that variation in the genetic and molecular programs unfolding within cells during development can cause disease. For example, most Mendelian disorders have a major developmental component (*4*). More-common and often devastating developmental conditions to which genetic factors substantially contribute include congenital heart defects, other birth defects, intellectual disabilities, and autism (*5*).

Several challenges have historically limited the study of developing human tissues at the molecular level. First, access to human embryonic and fetal tissues is limited. Second, even when available, the tissues are usually fixed and nucleic acids are degraded. Third, until recently, most molecular studies of complex tissues have been confounded by cell type heterogeneity. For these reasons, contemporary knowledge of the molecular basis of in vivo human development mostly derives from a combination of human genetics (in particular, of Mendelian disorders), in vivo investigations of model organisms (in particular, of the mouse), and in vitro studies of differentiating human cell lines (in particular, of embryonic or induced pluripotent stem cells), rather than from direct investigations of developing human tissues.

A reference human cell atlas based on developing tissues could serve as the foundation for a systematic effort to better understand the molecular and cellular events that give rise to all rare and common disorders of development, which collectively account for a major proportion of pediatric morbidity and mortality (*6*, *7*). Furthermore, although pioneering cell atlases have already been reported for many adult human organs (*8*, *9*), developing tissues may provide better opportunities to study the in vivo emergence and differentiation of human cell types. Relative to embryonic and fetal tissues, adult tissues are dominated by differentiated cells, and many cell states are not represented. By better resolving cell types and their trajectories, single-cell atlases generated from developing tissues could broadly inform our basic understanding of human biology as well as strategies for cell reprogramming and cell therapy.

As one step toward a comprehensive cell atlas of human development (*10*), we set out to generate single-cell atlases of both gene expression and chromatin accessibility using diverse human tissues obtained during mid-gestation (DESCARTES, Developmental Single Cell Atlas of gene Regulation and Expression; descartes.brotmanbaty.org). For gene expression, we applied three-level single-cell combinatorial indexing (sci-RNA-seq3) (*11*) to 121 fetal tissues representing 15 organs, altogether profiling gene expression in 5 million cells (Fig. 1A and table S1). We also measured chromatin accessibility in 1.6 million cells from the same organs using an overlapping set of samples (*12*). The profiled organs span diverse systems; however, some systems were not accessible—bone marrow, bone, gonads, and skin are notably absent.

Tissues were obtained from 28 fetuses ranging from 72 to 129 days in estimated post-conceptual age. We applied a method for extracting nuclei directly from cryopreserved tissues that works across a variety of tissue types and produces homogenates suitable for both sci-RNA-seq3 and sci-ATAC-seq3 (single-cell combinatorial indexing assay for transposase-accessible chromatin with high-throughput sequencing) (*12*). For most organs, extracted nuclei were fixed with paraformaldehyde. For renal and digestive organs where ribonucleases (RNases) and proteases are abundant, we used fixed cells rather than nuclei, which increased cell and mRNA recovery (*13*). For each experiment, nuclei or cells from a given tissue were deposited to different wells, such that the first index of sci-RNA-seq3 protocol also identified the source. As a batch control for experiments on nuclei, we spiked a mixture of human HEK293T and mouse NIH/3T3 nuclei, or nuclei from a common sentinel tissue (trisomy 21 cerebrum), into one or several wells. As a batch control for experiments on cells, we spiked cells derived from a tissue (pancreas) into one or several wells.

We sequenced sci-RNA-seq3 libraries from seven experiments across seven Illumina NovaSeq 6000 sequencer runs, altogether generating 68.6 billion raw reads. Processing data as previously described (*11*), we recovered 4,979,593 single-cell gene expression profiles [unique molecular identifier (UMI) > 250] [see files S1 to S3 at the Gene Expression Omnibus (GEO) (accession no. GSE156793)]. Single-cell transcriptomes from human-mouse control wells were overwhelmingly species coherent (~5% collisions) (fig. S1A). Uniform manifold approximation and projection (UMAP) (*14*) of nuclei or cells from the sentinel tissues indicated that cell type differences dominated over interexperimental batch effects (fig. S1, B and C). Integrated analysis (*15*) of nuclei and cells from the common pancreatic tissue also resulted in highly overlapping distributions (fig. S1D).

We profiled a median of 72,241 cells or nuclei per organ [Fig. 1A; maximum, 2,005,512 (cerebrum); minimum, 12,611 (thymus)]. Despite shallow sequencing (~14,000 raw reads per cell) relative to other large-scale single-cell RNA sequencing (scRNA-seq) atlases (*16*–*19*), we recovered a comparable number of UMIs per cell or nucleus (median 863 UMIs and 524 genes, not including cultured cells; fig. S1E). As expected, nuclei exhibited a higher proportion of UMIs mapping to introns than cells (56% for nuclei; 45% for cells; $P < 2.2 \times 10^{-16}$, two-sided Wilcoxon rank sum test). We

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA. [3]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [4]Department of Pediatrics, University of Rochester Medical Center, Rochester, NY, USA. [5]Center for Data Visualization, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [6]Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, USA. [7]Illumina Inc., San Diego, CA, USA. [8]Human Molecular Genomics Group, Max Planck Institute for Molecular Genetics, Berlin, Germany. [9]Institute of Human Genetics, University of Lübeck, Lübeck, Germany. [10]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA. [11]Howard Hughes Medical Institute, Seattle, WA, USA.
*Present address: Laboratory of Single-Cell Genomics and Population Dynamics, The Rockefeller University, New York, NY, USA.
†Corresponding author. Email: coletrap@uw.edu (C.T.); shendure@uw.edu (J.S.)

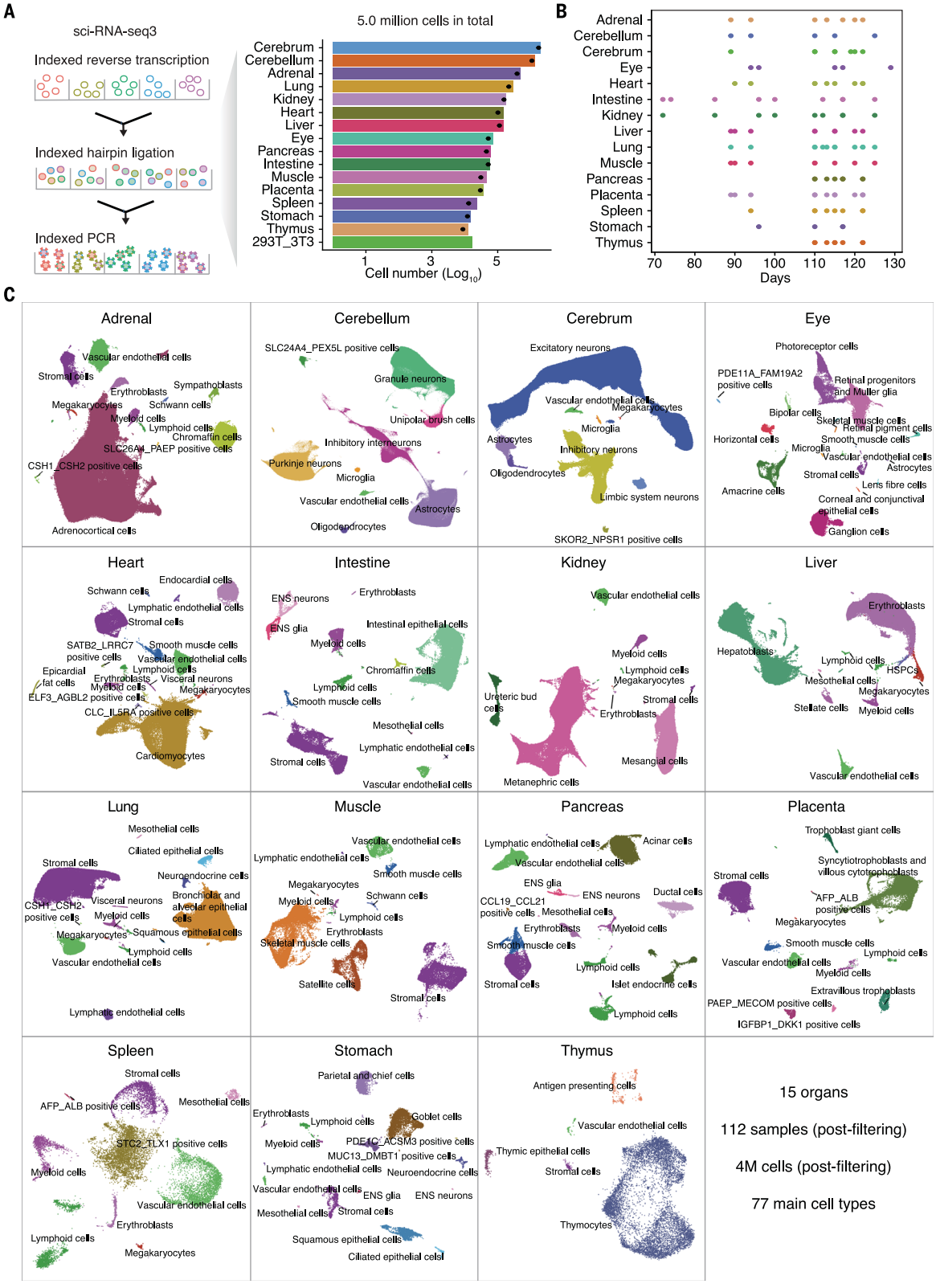


Fig. 1. Data generation and identifying cell types across 15 human organs. (**A**) Project workflow (left) and bar plot (right) showing the number of cells profiled per organ on a $\log_{10}$ scale. Dots indicate the number of cells remaining for downstream analysis after quality control (QC) filtering procedures. PCR, polymerase chain reaction. (**B**) Bar plot showing the distribution of estimated postconceptual ages for tissue samples corresponding to each organ. (**C**) After filtering against low-quality cells and doublet-enriched clusters, 4 million single-cell gene expression profiles were subjected to UMAP visualization and Louvain clustering with Monocle 3 on a per-organ basis. Clusters were initially annotated on a per-organ basis as well, utilizing recent organ-specific cell atlas efforts, which yielded 172 main cell types (colors and labels). Because many cell type annotations appear in multiple organs (e.g., vascular endothelial cells), we consolidated these to 77 main cell types.

henceforth use the word cells to refer to both cells and nuclei, unless otherwise stated.

Tissues were readily identified as deriving from a male (n = 14) or a female (n = 14) by sex-specific gene expression (fig. S1F). Each of the 15 organs was represented by multiple samples (median 8) that included at least two of each sex (fig. S1G) and a range of estimated postconceptual ages (Fig. 1B). Pseudobulk transcriptomes clustered by organ rather than individual or experiment [fig. S1H; files S4 and S5 at GEO (GSE156793)]. About half of the expressed, protein-coding transcripts were differentially expressed across pseudobulk transcriptomes [11,766 of 20,033; false discovery rate (FDR) of 5%; table S2].

We applied Scrublet (*20*) to detect 6.4% likely doublet cells, which corresponded to a doublet estimate of 12.6% including both within-cluster and between-cluster doublets (fig. S1I). We then applied a scalable strategy that we had previously developed (*11*) to remove low-quality cells, doublet-enriched clusters, and the spiked-in HEK293T and NIH/3T3 cells. All analyses below focus on the 4,062,980 human single-cell gene expression profiles derived from 112 fetal tissue samples that remained after this filtering step.

Identification and annotation of 77 main cell types

Using Monocle 3 (*11*), we subjected single-cell gene expression profiles to UMAP visualization and Louvain clustering on a per-organ basis. Altogether, we initially identified and annotated 172 cell types on the basis of cell type–specific marker gene expression (*16*, *21*–*84*) [Fig. 1C and table S3; files S6 and S7 at GEO (GSE156793)]. After collapsing common annotations across tissues, these reduced to 77 main cell types, 54 of which were observed in only a single organ (e.g., Purkinje neurons in the cerebellum) and 23 of which were observed in multiple organs (e.g., vascular endothelial cells in every organ). There were 15 cell types that we were unable to annotate during our manual, organ-by-organ review (the subset named by a pair of markers in Fig. 1C); these are discussed further below and in (*85*). Each of these 77 main cell types was represented by a median of 4829 cells, ranging from 1,258,818 cells (excitatory neurons in the cerebrum) to only 68 cells (*SLC26A4*- and *PAEP*-positive cells in the adrenal gland) (fig. S2A). Each main cell type was observed in multiple individuals (median 9; fig. S2B). We recovered nearly all major cell types identified by previous atlasing efforts directed at the same organs, despite differences with respect to species, stage of development, and technology (*16*, *23*, *28*, *33*, *35*, *51*, *69*, *72*, *86*–*88*). We identified a median of 12 main cell types per organ, ranging from 5 (thymus) to 16 (eye, heart, and stomach). We did not observe a correlation between the number of profiled cells and the number of identified cell types (Spearman $\rho = -0.10$, $P = 0.74$).

On average, we identified 11 marker genes per main cell type (minimum, 0; maximum, 294; defined as differentially expressed genes with at least a fivefold difference between first and second ranked cell type with respect to expression; FDR of 5%; fig. S2C and table S4). There were several cell types that lacked marker genes at this threshold because of highly related cell types in other organs (e.g., enteric glia versus Schwann cells). For this reason, we also report sets of within-tissue marker genes, determined by the same procedure but on an organ-by-organ basis (average 147 markers per cell type; minimum, 12; maximum, 778; fig. S2D and table S5). An interactive website facilitates the exploration of these data by tissue, cell type, or gene (descartes.brotmanbaty.org).

Although canonical markers were generally observed and were critical for our annotation process, to our knowledge, most of the observed markers have not been identified in prior studies. For example, *OLR1*, *SIGLEC10*, and noncoding RNA *RP11-480C22.1* are among the strongest markers of microglia, along with more-established microglial markers such as *CLEC7A* (*89*), *TLR7* (*90*), and *CCL3* (*91*). As anticipated, given that these tissues are undergoing development, many of the 77 main cell types include states progressing from precursors to one or several terminally differentiated cell types. For example, cerebral excitatory neurons exhibited a continuous trajectory from *PAX6*+ neuronal progenitors to *NEUROD6*+ differentiating neurons (*92*) to *SLC17A7*+ mature neurons (*93*) (fig. S2, E and F). In the liver, hepatic progenitors (*DLK1*+, *KRT8*+, and *KRT18*+) (*94*, *95*) exhibited a continuous trajectory to functional hepatoblasts (*SLC22A25*+, *ACSS2*+, and *ASS1*+) (fig. S2, G and H) (*96*–*98*). In contrast with mouse organogenesis—wherein the maturation of the transcriptional program is tightly coupled to developmental time (*11*)—cell state trajectories were inconsistently correlated with estimated postconceptual ages in these data (fig. S2, I and J). A potential explanation for this is that gene expression is markedly more dynamic during embryonic than during fetal development. However, it is also possible that inaccuracies in the estimated postconceptual ages confound our resolution.

In addition to these manual annotations of cell types, we also generated semiautomated classifiers for each organ using Garnett (*99*). The Garnett classifiers were generated agnostic of previous clustering, with marker genes separately compiled from the literature (*99*). Classifications by Garnett were concordant with manual classifications (fig. S3A). Using the Garnett models trained on these data, we were able to accurately classify cell types from other single-cell datasets, including data generated with different methods as well as those from adult organs. When we applied the classifier for pancreas to inDrop scRNA-seq data (*100*), Garnett correctly annotated 82% of the cells (cluster-extended; 11% incorrect, 8% unclassified) (fig. S3B). These models can broadly be used for the automated cell type classification of single-cell data from diverse organs (fig. S3C; descartes.brotmanbaty.org).

We next evaluated the specificity of our main cell types by intradataset cross-validation with a support vector machine (SVM) classifier (*101*). In this framework, high cross-validation precision and recall values indicate that cells derived from a given cluster can robustly be reassigned to that cluster; we thus use high F1 scores as a proxy for identifying cell clusters as valid types, at least in the setting of the tissue in which they were identified. We first evaluated this approach on the kidney. As expected, annotated kidney cell types have much higher specificity scores (median 0.99) than control cell types, in which cell labels are permuted before cross-validation (median 0.17) [Fig. 2A (leftmost panel only), Fig. 2B (left panel only), fig. S4A, and table S3].

We then applied this approach to cells from each organ. Once again, annotated main cell types exhibited much higher specificity scores than permuted cell types (Fig. 2C and fig. S4B; median 0.99 versus 0.10; $P < 2.2 \times 10^{-16}$, two-sided Wilcoxon rank sum test). Despite smaller numbers of cells, most of the 15 initially unannotated cell types also exhibited high specificity scores (median 0.98). The exceptions are probably better described as subtypes of other cell types [discussed further below and in (*85*)]. We also applied this method to the consolidated set of 77 main cell types (i.e., rather than organ-by-organ) with similar results (fig. S4C).

Automated preliminary annotation of cell subtypes

To identify cell subtypes, we performed unsupervised clustering on main cell types with >1000 cells in any given tissue. For each main cell type in each tissue, we first applied batch correction (*102*) followed by dimensionality reduction and Louvain clustering (Fig. 2A). After merging clusters that were not readily distinguishable by the intradataset cross-validation procedure described above, a total of 657 cell subtypes were identified across the 15 tissues, with a median of 824 cells in each. All subtypes were composed of cells contributed by at least two individuals (median 7). Unsurprisingly, given the procedure used for merging clusters, these subtypes have higher specificity scores than permuted controls (median 0.77 versus 0.13; $P < 2.2 \times 10^{-16}$, two-sided Wilcoxon rank sum test; Fig. 2C).

We next sought to leverage existing mouse cell atlases to annotate these human subtypes

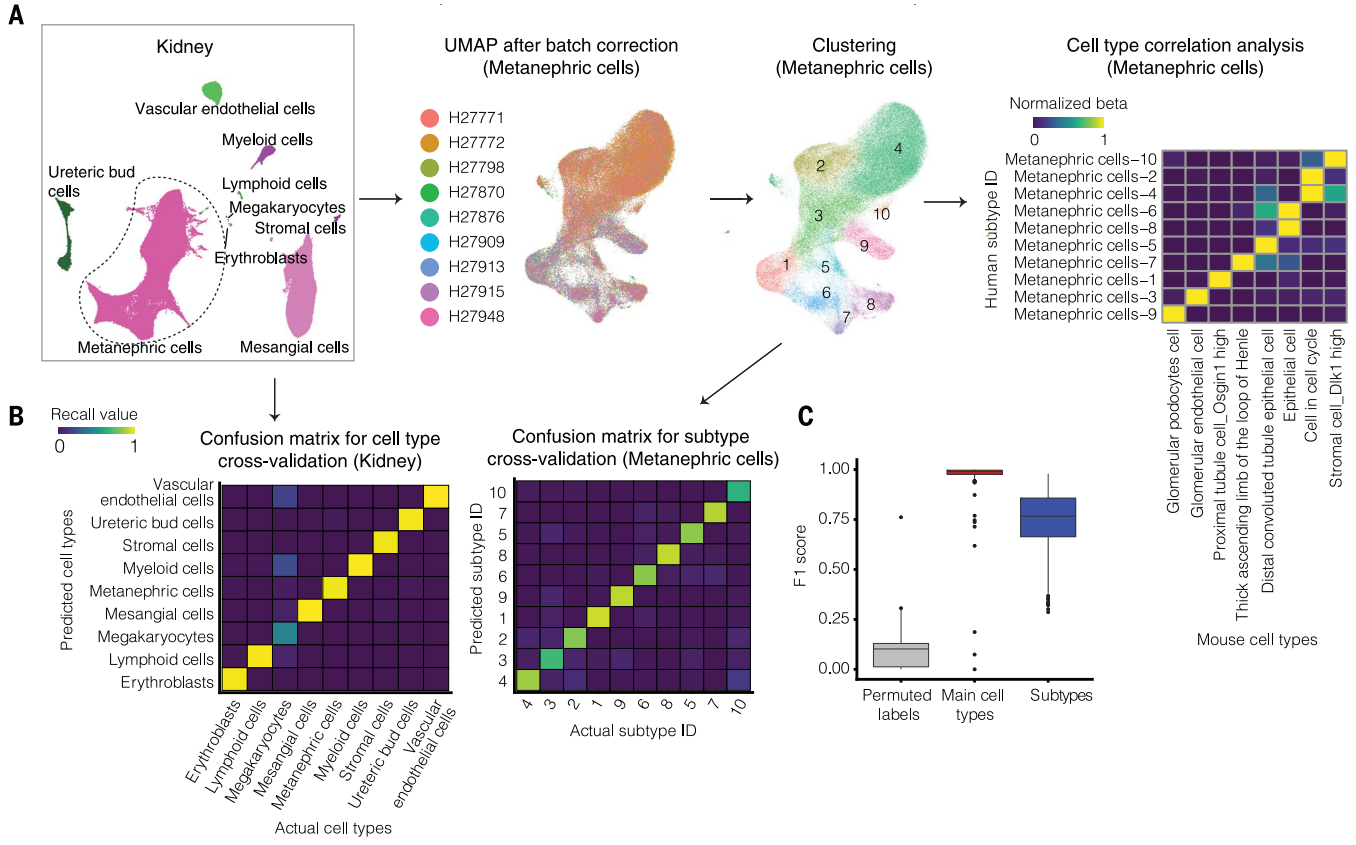


Fig. 2. Identification of cell subtypes. (**A**) Pipeline for cell subtype identification. Briefly, on a tissue-by-tissue basis, we subjected each main cell type with >1000 cells to batch correction (*102*), UMAP visualization, and Louvain clustering. Clusters with similar transcriptomes were merged by an automated procedure. Briefly, we applied an intradataset cross-validation approach (*101*) to evaluate their specificity and iteratively merged similar clusters. We then compared putative human cell subtypes identified in our data (rows) against annotated mouse cell types from the corresponding tissues (*16*) (columns) by cell type correlation analysis. Colors correspond to beta values, normalized by the maximum beta value per row. All MCA cell types with a beta of a matched human cell type >0.01—i.e., also the maximum beta for that human cell type—are shown for the kidney metanephric cells. (**B**) Confusion matrix for intradataset cell type cross-validation with an SVM classifier for main cell types (left) and metanephric subtypes (right) in the kidney. In total, 2000 cells (or all cells for cell types with <2000 cells profiled) are randomly sampled for each cell type or subtype before cross-validation analysis. (**C**) Box plot showing the cell specificity score (F1 score) distribution for permuted controls, main cell types, and subtypes from intradataset cross-validation.

in an automated fashion. With a cell type cross-matching method that we had previously developed (*11*), we could match 605 of 606 (99%) human cell subtypes to at least one cell type in corresponding fetal and/or adult tissues from the mouse cell atlas (MCA) (*16*) (specificity score beta > 0.01, the same threshold that we used to align against MCA previously; 51 adrenal subtypes were excluded because corresponding MCA tissue was not available) (table S6 and figs. S5 to S8). Additionally, 77 of 148 (52%) cerebral or cerebellar subtypes matched to at least one adult cell type from the mouse brain cell atlas (MBCA) (fig. S9) (*50*).

Despite the species difference, many human cell subtypes matched 1:1 with mouse cell types. For example, diverse epithelial subtypes in the human kidney matched 1:1 with annotated MCA cell types (Fig. 2A), and diverse neuronal subtypes in the human cerebrum matched 1:1 with annotated MBCA cell types (fig. S9). Notably, although there were many sets of human subtypes that matched a single MCA or MBCA cell type (e.g., hepatoblasts in fig. S5 and oligodendrocytes in fig. S9), these likely reflect bonafide heterogeneity as evidenced by their specificity scores (Fig. 2C). Additional work is necessary to annotate such subtypes with greater granularity.

Integration across tissues and investigation of initially unannotated cell types

We next sought to integrate data and compare cell types across all 15 organs. To mitigate the effects of gross differences in sampling, we randomly sampled 5000 cells per cell type per organ (or in cases where <5000 cells of a given cell type were represented in a given organ, all cells were taken), and we performed UMAP visualization (Fig. 3A and fig. S10A). As expected, cell types represented in multiple organs, as well as developmentally related cell types, tended to colocalize. Many surface proteins (4565 of 5480), secreted proteins (2491 of 2933), transcription factors (1715 of 1984), and noncoding RNAs (3130 of 10,695) were differentially expressed across the 77 main cell types (FDR of 0.05; Fig. 3B and table S4; descartes.brotmanbaty.org). The expression patterns of noncoding RNAs were notably sufficient to separate cell types into developmentally coherent groups (fig. S10, B and C).

As noted above, there were 15 cell types that we were unable to annotate during our manual, organ-by-organ review (the subset named by pairs of markers in Fig. 1C). To shed light on these, we examined their distribution in the global UMAP (Fig. 3A), whether they matched to annotated cell types in MCA or MBCA (figs. S5 to S9), their distribution across tissues derived from different individuals (fig. S11A), and their potential for maternal origin (fig. S11B).

These further analyses enabled us to annotate 8 of the 15 cell types (*85*). For example, rare *CSH1*- and *CSH2*-positive cells in the lung and adrenal gland (two of the most deeply profiled organs)

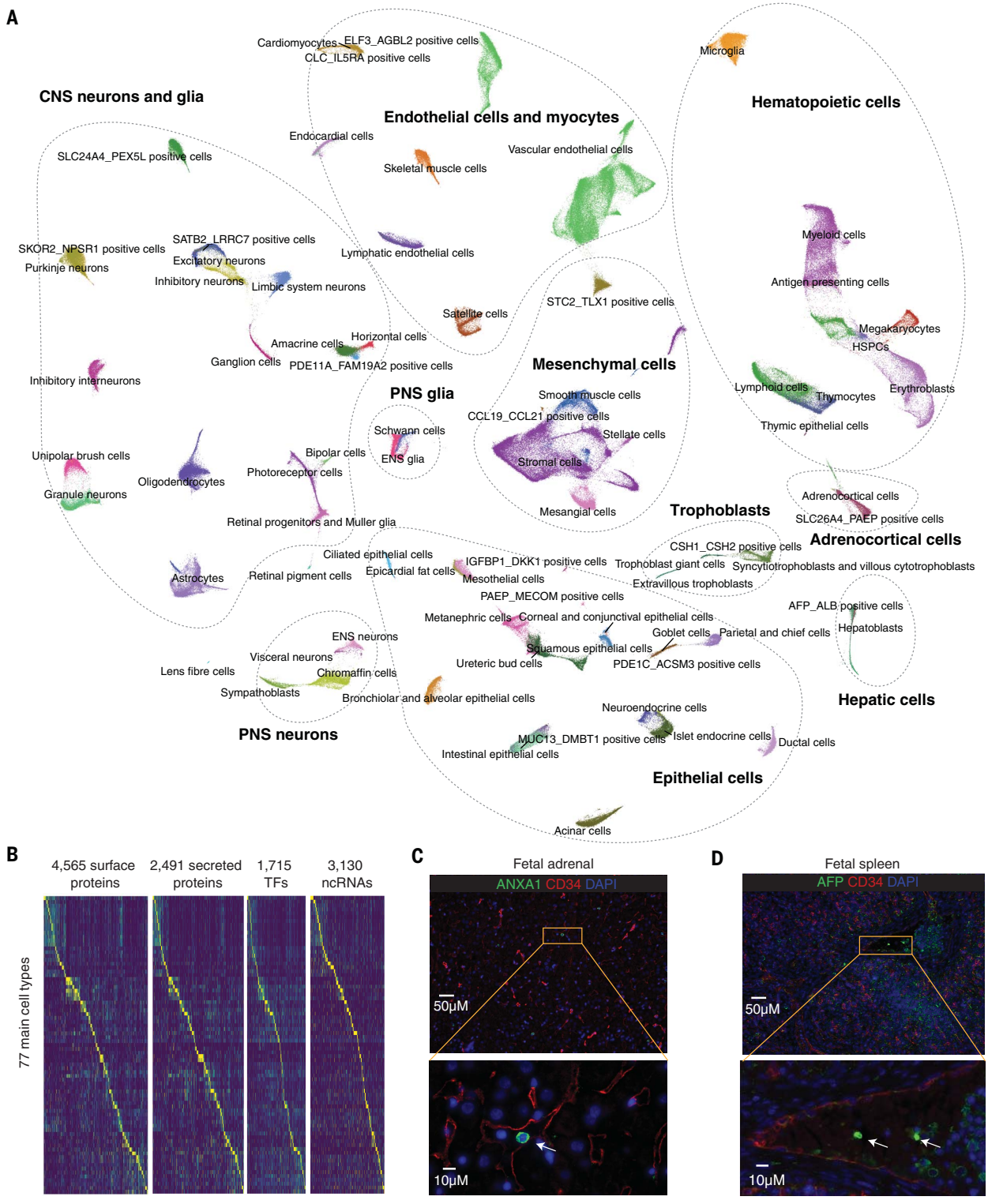


Fig. 3. Integrated visualization of cell types across all profiled tissues. (**A**) From each organ, we sampled 5000 cells from each cell type (or all cells for cell types with <5000 cells in a given organ). These were subjected to UMAP visualization on the basis of the top differentially expressed genes across cell types within each organ. Here, they are colored by cell type labels, with colors as in Fig. 1C. In fig. S10A, the same UMAP visualization is colored by tissue of origin. (**B**) Heatmap showing the relative expression of surface and secreted protein-coding genes, noncoding RNAs (ncRNAs), and TFs (columns) in 77 main cell types (rows). UMI counts for genes are scaled for library size, log-transformed, and then mapped to *Z* scores and capped to [0, 3]. (**C** and **D**) Representative fluorescence microscopy images of (C) human fetal adrenal or (D) spleen tissue, staining for (C) endothelium (CD34+), *CSH1*-, and *CSH2*-positive cells (ANXA1+; labeled by arrowhead) or (D) *AFP*- and *ALB*-positive cells (AFP+ is indicated with arrows). Nuclei are stained with blue 4′,6-diamidino-2-phenylindole (DAPI). Bottom panels correspond to inset zooms. Scale bars, 50 μm (top) and 10 μm (bottom).

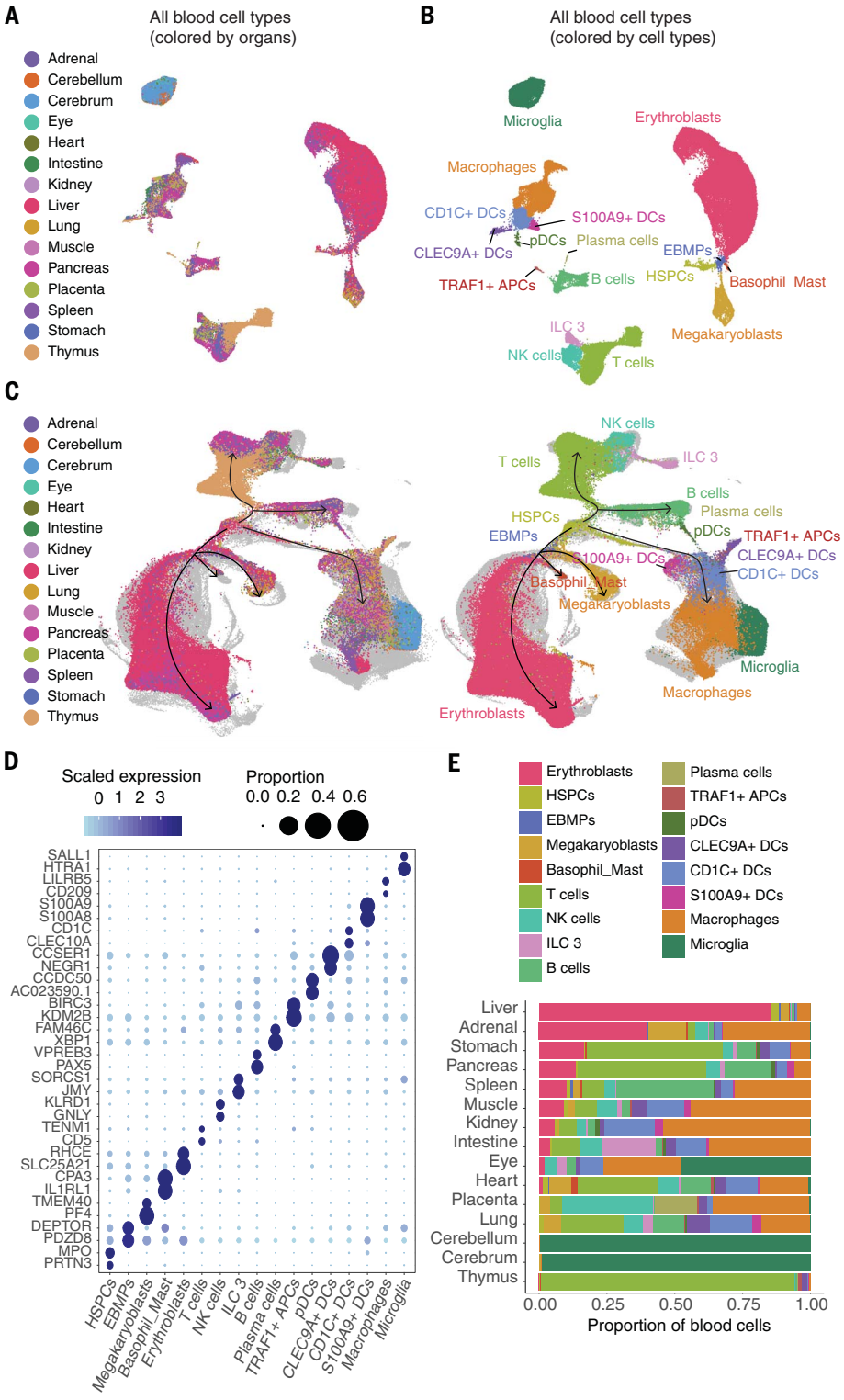


Fig. 4. Identification and characterization of blood cell subtypes and developmental trajectories. (**A** and **B**) UMAP visualization and marker-based annotation of blood cell types colored by organ type (A) and cell type (B). (**C**) UMAP visualization of blood cells, integrating across all profiled organs of this study and an scRNA-seq atlas of blood cells from human fetal liver (*108*). Cells from (*108*) are colored in light gray, and cells from our study are colored by tissue of origin (left) or blood cell types (right). Black arrows indicate inferred cell state transition directions from HSPCs to all main blood lineages. (**D**) Dot plot showing expression of two selected marker genes per cell type. The size of the dot encodes the percentage of cells within a cell type in which that marker was detected, and its color encodes the average expression level. (**E**) Bar plot showing the estimated fraction of cells per organ derived from each of the 17 annotated blood cell types.

are highly similar to placental trophoblasts—e.g., expressing high levels of placental lactogen, chorionic gonadotropin, and aromatase (Fig. 3A) (*85*). *AFP*- and *ALB*-positive cells in the placenta and spleen resemble hepatoblasts—e.g., expressing high levels of serum albumin, alpha fetoprotein, and apolipoproteins (Fig. 3A) (*85*) [at least in the placenta, similar hepatoblast-like *AFP*- and *ALB*-positive cells were observed in the mouse (fig. S5)]. Follow-up immunostaining studies supported the presence of these trophoblast-like and hepatoblast-like cells in the adrenal gland and spleen, respectively (Fig. 3, C and D, and fig. S12). Given that these cell types are rarely but recurrently observed in several organs, they potentially correspond to circulating trophoblasts and circulating hepatoblasts.

In males, both *IGFBP1*- and *DKK1*-positive as well as *PAEP*- and *MECOM*-positive cells in the placenta expressed appreciable levels of *XIST* or *TSIX* (fig. S12B); on further review of markers, these correspond to maternal decidualized stromal cells and maternal endometrial epithelial cells, respectively. This conclusion is supported by maternal genotypes in the corresponding cell types in chromatin accessibility data (*12*).

Several additional cell types were annotated through strong matches to MCA or MBCA (fig. S13) or through their position in the global UMAP coupled with additional literature review (Fig. 3A) (*85*); these include *STC2*- and *TLX1*-positive cells, which are abundant in the spleen and express genes associated with mesenchymal precursor or stem cells (*103*–*105*). Of the remaining seven initially unannotated cell types, four would likely better be classified as subtypes (and correspondingly, these tended to have lower specificity scores), and three have high specificity scores but remain ambiguous (*85*).

Characterization of blood lineage development across organs

The nature of this dataset creates an opportunity to systematically investigate organ-specific differences in gene expression within broadly distributed cell types—for example, blood cells. We reclustered 103,766 cells, derived from all 15 organs, that corresponded to hematopoietic cell types (Fig. 4A). We then performed Louvain clustering and further annotated fine-grained blood cell types, in some cases identifying very rare cell types (Fig. 4B). For example, myeloid cells separate into microglia, macrophages, and diverse dendritic cell subtypes [*CD1C*+, *S100A9*+, *CLEC9A*+, and plasmacytoid dendritic cells (pDCs)] (*106*). The microglial cluster primarily derives from brain tissues, and it is well separated from macrophages, which is consistent with their distinct developmental trajectories (*107*). Lymphoid cells clustered into several groups, including B cells, natural killer (NK) cells, ILC 3 cells, and T cells, the latter of which

includes the thymopoiesis trajectory. We also recovered very rare cell types such as plasma cells (139 cells, mostly in placenta and making up 0.1% of all blood cells or 0.003% of the full dataset) and *TRAF1*+ antigen-presenting cells (APCs) (189 cells, mostly in thymus and heart and making up 0.2% of all blood cells or 0.005% of the full dataset).

To validate these annotations, we integrated fetal blood cells from all organs with an scRNA-seq atlas of blood cells from the fetal liver (*108*) (Fig. 4C, left, and fig. S14A). Despite different methods, corresponding cell types from two datasets were highly overlapping; this was also the case upon integration analysis with another scRNA-seq dataset of 1231 human embryonic blood cells (*109*) (fig. S14B). Notably, some extremely rare cell types identified through CD45+ fluorescence-activated cell sorting (FACS) enrichment (e.g., *VCAM1*+ EI macrophages, monocyte precursors, and neutrophil-myeloid progenitors) were not annotated in our data. On the other hand, we captured fetal blood cells derived from tissues other than the liver—e.g., microglia in the brain and T and B cells in the thymus and spleen, respectively. Furthermore, as they span multiple organs, we are better able to capture cell state transition paths from hematopoietic stem and progenitor cells (HSPCs) to lymphoid cells than a single-organ study (Fig. 4C, right).

Although gene expression markers for different immune cell types have been extensively studied, these may be limited by their definition via a restricted set of organs or cell types. Here, we find that many conventional immune cell markers were expressed in multiple cell types. For example, conventional markers for T cells (*110*–*112*) were also expressed in macrophages and dendritic cells (*CD4*) or NK cells (*CD8A*), consistent with other studies (*113*) (fig. S14C). We computed pan-organ cell type–specific markers across 14 blood cell types (Fig. 4D and table S7). From this we observed that T cells specifically expressed *CD8B* and *CD5* (*114*) as expected, but also *TENM1* (Fig. 4D and fig. S14C). ILC 3 cells, whose annotation was determined on the basis of their expression of *RORC* (*115*) and *KIT* (*116*), were more specifically marked by *SORCS1* and *JMY* (Fig. 4D and fig. S14C). These and other markers identified by pan-organ analysis may be useful for labeling and purifying specific blood cell types.

As expected, different organs showed varying proportions of blood cells (Fig. 4E). For example, the liver contained the highest proportion of erythroblasts, consistent with its role as the primary site of fetal erythropoiesis (*117*), whereas T cells were enriched in the thymus and B cells in the spleen. Nearly all blood cells recovered from the cerebellum and cerebrum were microglia. The tissue distribution of ILC 3 cells as well as subtypes of dendritic cells was captured as well (Fig. 4E and fig. S14D). Pan-organ analysis also enabled the identification of rare cell populations in specific organs. We identified rare HSPCs in the liver but also rare cells that are transcriptionally similar to HSPCs in the lung, spleen, thymus, heart, intestine, adrenal gland, and other organs (fig. S15). Subclustering analyses showed that HSPCs outside of the liver, as well as a subset of liver HSPCs, expressed differentiation markers such as *LYZ* (*118*), *ACTG1* (*119*), and *ANK1* (*120*), whereas most liver HSPCs expressed *MECOM* and *NRIP1*, both of which are required for the maintenance and function of normal quiescent HSPCs (*121*, *122*) (fig. S15).

Focusing on erythropoiesis, we observed a continuous trajectory from HSPCs to an intermediate cell type, erythroid-basophil-megakaryocyte biased progenitors (EBMPs), which then split into erythroid, basophilic, and megakaryocytic trajectories (Fig. 5A and table S8), consistent with a recent study of the mouse fetal liver (*123*, *124*). This consistency was despite differences in species (human versus mouse), techniques (sci-RNA-seq3 versus 10x Genomics), and tissues (pan-organ versus liver only). With unsupervised clustering and adopting terminology from that study (*123*), we further partitioned the continuum of erythroid states into three stages: early erythroid progenitors (EEPs) (marked by *SLC16A9* and *FAM178B*), committed erythroid progenitors (CEPs) (marked by *KIF18B* and *KIF15*), and cells in the erythroid terminal differentiation state (ETDs) (marked by *TMCC2* and *HBB*) (Fig. 5B). Early and late stages of megakaryocytic cells were also readily identified (Fig. 5, A and B).

As expected, given their established role in fetal erythropoiesis, a portion of blood cells in the liver and spleen corresponded to EEPs, CEPs, and megakaryocyte progenitors (*125*). Notably, we also observed EEPs, CEPs, and megakaryocyte progenitors in the adrenal gland in every sample studied (Fig. 5C and fig. S16A). Because we do not observe cell types that are more common in the liver and spleen, trivial contamination during recovery of the adrenal glands is an unlikely explanation. Although occasional islands of extramedullary hematopoiesis have been observed in the adrenal glands of human embryos (*126*, *127*), the consistency across individuals led us to further investigate whether the adrenal glands may serve as a normal site of erythropoiesis in mammals. Immunohistochemical analysis of human fetal adrenal tissues showed nucleated GYPA+ cells outside CD34+ blood vessels (Fig. 5D and fig. S16B). We further used imaging flow cytometry to visualize and enumerate maturing erythroid precursors and enucleated erythrocytes (*128*) in the perinatal period of the mouse. Approximately 8% of viable dissociated cells from the adrenal gland consisted of maturing erythroblasts, compared with 0.2% of viable dissociated cells in the kidney (Fig. 5E). Also consistent with the adrenal gland being a site of ongoing erythropoiesis, its distribution of immature to mature erythroblasts matched closely with that of the bone marrow of adult mice (Fig. 5, E and F).

Macrophages were even more widely distributed. We collated all macrophages, together with microglia from the brain, and subjected them to UMAP visualization and Louvain clustering, independent of other cell types (Fig. 5, G and H; fig. S16C; and table S9). Notably, microglia were divided into three subclusters, one of which, marked by *IL1B* and *TNFRSF10D*, likely represents activated microglia expressing proinflammatory cytokines involved in the normal development of the nervous system (*129*, *130*). The other microglial clusters were marked by expression of *TMEM119* and *CX3CR1* (*131*) (more common in the cerebrum) or *PTPRG* and *CDC14B* (*132*) (more common in the cerebellum).

The macrophages outside the brain clustered into three major groups (Fig. 5, G and H; fig. S16C; and table S9): (i) antigen-presenting macrophages, found mostly in gastrointestinal (GI) tract organs (intestine and stomach) and marked by high expression of antigen-presenting (e.g., *HLA-DPB1* and *HLA-DQA1*) and inflammatory activation genes [e.g., *AHR* (*133*)]; (ii) perivascular macrophages, found in most organs, with specific expression of markers such as *F13A1* (*134*) and *COLEC12* (*135*), as well as markers such as *RNASE1* and *LYVE1*; and (iii) phagocytic macrophages, enriched in the liver, spleen, and adrenal gland (Fig. 5I), with specific expression of markers such as *CD5L* (*136*), *TIMD4* (*137*), and *VCAM1* (*138*). Phagocytic macrophages are critical for removing the pyrenocytes (the so-called extruded nucleus) after enucleation of late-stage erythroblasts to form reticulocytes; their observation in the adrenal gland is consistent with its aforementioned potential role as an additional site of normal fetal erythropoiesis. Below, we leverage integration with a mouse atlas of organogenesis (*11*) to investigate the conserved program of blood cell specification and developmental origins of microglia and macrophages.

Characterization of endothelial and epithelial cells across organs

As a second analysis of a single class of cells across many organs, we reclustered 89,291 endothelial cells (ECs) that correspond to vascular endothelium (VECs), lymphatic endothelium (LECs), or endocardium. These three groups readily separated from one another, and VECs further clustered, at least to some degree, by organ (fig. S17, A to C). That organ-specific differences are more readily detected than differences between arteries, capillaries, and veins is consistent with previous cell atlases of the adult mouse (*16*, *28*). We performed

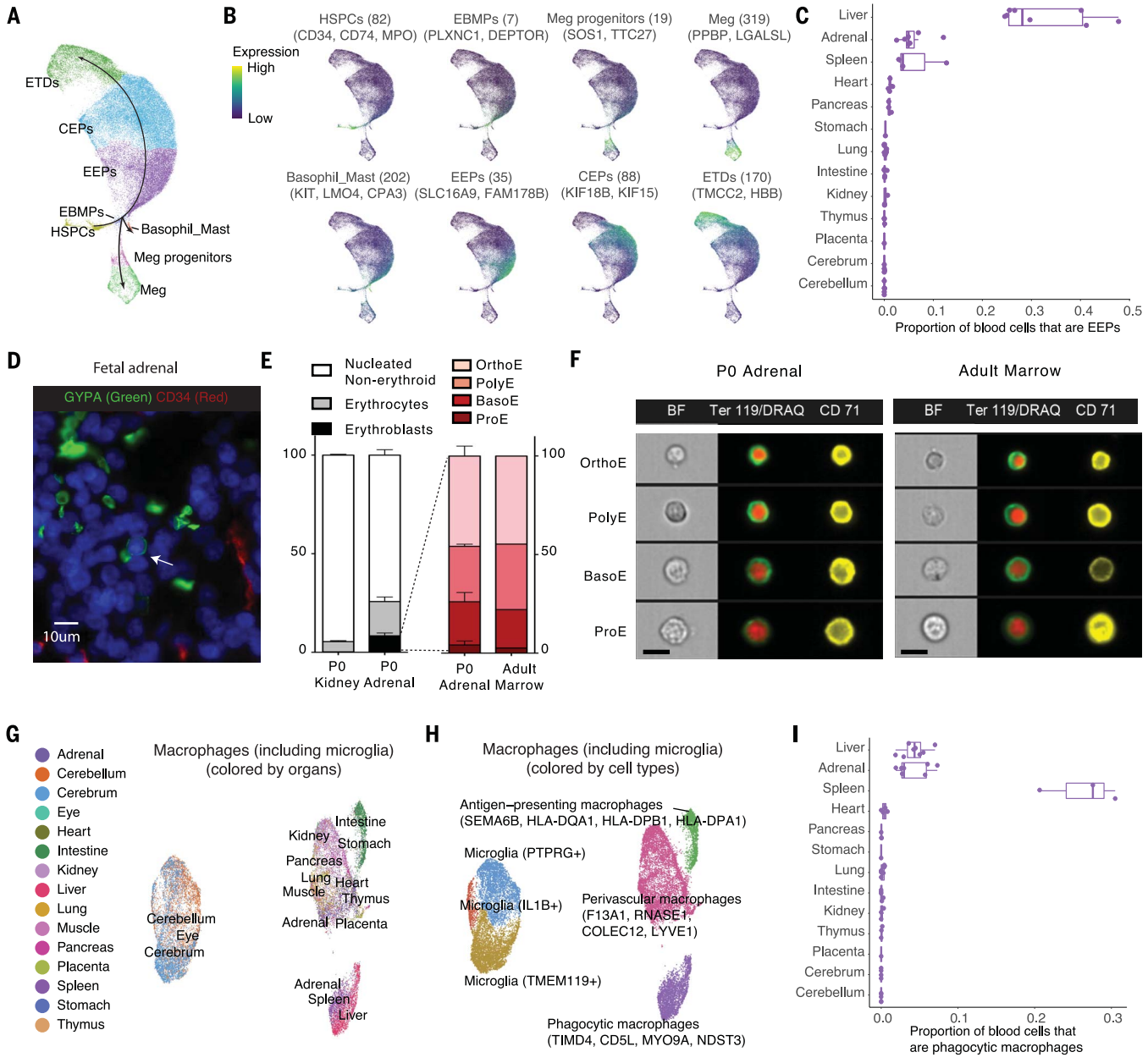


Fig. 5. Identification and characterization of erythropoiesis and macrophage differentiation in adrenal gland. (**A**) Zoomed view of the erythropoiesis trajectory portion of Fig. 4B, colored by erythroid or megakaryocyte subtype. Black arrows show trajectory directionalities defined by (*123*). (**B**) Plots similar to (A), colored by the normalized expression of cell type–specific genes (FDR of 0.05 and more than twofold expression difference between first and second ranked cell type), with the number of cell type–specific genes used and names of the top few genes shown. UMI counts for these genes are scaled for library size, log-transformed, aggregated, and then mapped to *Z* scores. (**C**) Point and box plot showing the proportion of blood cells that are EEPs for individual samples of different organs. Samples with low recovery of blood cells (≤200) are excluded. (**D**) Representative fluorescence microscopy of human fetal adrenal tissue, staining for endothelium (CD34+) and erythroblasts (nucleated and GYPA+); nuclei stained with blue DAPI. The arrow indicates a GYPA+ erythroblast outside a CD34+ blood vessel. Scale bars, 10 μm. (**E**) (Left) Percentage of dissociated kidney and adrenal glands from newborn (P0) mice composed of enucleated erythrocytes and maturing erythroblasts. (Right) Distribution of maturing erythroblasts (proerythroblasts, ProE; basophilic erythroblasts, BasoE; polychromatophilic erythroblasts, PolyE; and orthochromatic erythroblasts, OrthoE) in the adrenal gland at P0 and in adult bone marrow. Error bars represent means + SEM, $n = 3$. (**F**) Representative images of maturing erythroblasts in the P0 adrenal gland and the adult bone marrow. Scale bars, 10 μm. (**G** and **H**) UMAP visualization and marker-based annotation of macrophage subtypes colored by organ type (G) and subtype name (H). (**I**) Point and box plot showing the proportion of blood cells that are phagocytic macrophages for individual samples of different organs. Samples with low recovery of blood cells (≤200) are excluded.

an integrative analysis of ECs from human fetal tissues (this study) and mouse adult tissues (*139*) (fig. S17, D and E). Both human and mouse ECs were separated first by vascular versus lymphatic versus endocardial, and then by organ. VECs from the same tissue were generally clustered together, despite differences with respect to species, developmental stage, and technique. Conserved markers of organ-specific ECs were readily identified (fig. S17F) (*139*).

Differential gene expression analysis identified 700 markers that are specifically expressed

in a subset of ECs (FDR of 0.05 and more than twofold expression difference between first and second ranked cluster) (fig. S17G and table S10). About one-third of these encoded membrane proteins, many of which appeared to correspond to potential specialized functions (*12*, *140*–*142*). In agreement with observations in mice (*139*), brain ECs specifically expressed gene sets involved in amino acid transport ($q = 5.6 \times 10^{-10}$) and carboxylic acid transport ($q = 4.2 \times 10^{-8}$); lung ECs specifically expressed gene sets involved in adenosine 3′,5′-monophosphate (cAMP) ($q = 8.2 \times 10^{-3}$) and cyclic nucleotide ($q = 1.4 \times 10^{-2}$) catabolism, and vascular ECs from the GI tract, heart, and muscle specifically expressed gene sets involved in stem cell differentiation ($q = 3.7 \times 10^{-2}$). Potentially underlying these differences, human fetal ECs expressed distinct sets of transcription factors (TFs) (fig. S17H). For example, LECs specifically expressed *TBX1*, brain VECs specifically expressed *FOXQ1* and *FOXF2*, and liver VECs specifically expressed *DAB2*, all of which are consistent with observations in mice (*139*, *143*, *144*).

As a third analysis of a broadly distributed type of cell, we reclustered 282,262 epithelial cells, derived from all organs, and subjected these to UMAP visualization (fig. S18, A and B). Although some epithelial cell types were highly organ specific—e.g., acinar (pancreas) and alveolar cells (lung)—epithelial cells with similar functions generally clustered together (fig. S18C).

Within epithelial cells, two neuroendocrine cell clusters were identified (fig. S18C). The simpler of these corresponded to adrenal chromaffin cells and was marked by the specific expression of *HMX1* (*NKX-5-3*), a TF involved in sympathetic neuron diversification (*145*). The other cluster comprised neuroendocrine cells from multiple organs (stomach, intestine, pancreas, and lung) and was marked by specific expression of *NKX2-2*, a TF with a key role in pancreatic islet and enteroendocrine differentiation (*146*). We performed further analysis on the latter group, identifying five subsets (fig. S18, D to F): (i) pancreatic islet beta cells, marked by insulin expression; (ii) pancreatic islet alpha and gamma cells, marked by pancreatic polypeptide and glucagon expression; (iii) pancreatic islet delta cells, marked by somatostatin expression; (iv) pulmonary neuroendocrine cells (PNECs), marked by expression of *ASCL1* and *NKX2-1*, both TFs with key roles in specifying this lineage in the lung (*147*, *148*); and (v) enteroendocrine cells. Enteroendocrine cells further comprised several subsets, including *NEUROG*-expressing pancreatic islet epsilon progenitors (*149*, *150*), *TPH1*-expressing enterochromaffin cells in both the stomach and intestine (*151*), and gastrin- or cholecystokinin-expressing G, L, K, and I cells (*151*). Finally, we observed ghrelin-expressing enteroendocrine progenitors in the stomach and intestine (*150*, *152*), but also ghrelin-expressing endocrine cells in the developing lung (*153*) (fig. S18F). The diverse functions of neuroendocrine cells are closely linked with their secreted proteins; we identified 1086 secreted protein-coding genes differentially expressed across neuroendocrine cells (FDR of 0.05) (fig. S18G and table S11). For example, PNECs showed specific expression of trefoil factor 3, which is involved in mucosal protection and lung ciliated cell differentiation (*154*); gastrin-releasing peptide, which stimulates gastrin release from G cells in the stomach (*155*); and *SCGB3A2*, a surfactant associated with lung development (*156*).

As an illustrative example of how these data can be used to explore cell trajectories, we further investigated the path of epithelial cell diversification leading to renal tubule cells. Combining and reclustering ureteric bud and metanephric cells, we identified both progenitor and terminal renal epithelial cell types, with differentiation paths that are highly consistent with a recent study of the human fetal kidney (*157*) (fig. S19A). By differential gene expression analysis, we further identified TFs potentially regulating their specification (fig. S19B and table S12). For example, nephron progenitors in the metanephric trajectory specifically expressed high levels of mesenchyme and meis homeobox genes (*MEOX1*, *MEIS1*, and *MEIS2*) (*158*), whereas podocytes specifically expressed *MAFB* and *TCF21/POD1* (*159*, *160*). As another example, *HNF4A* was specifically expressed in proximal tubule cells—a mutation of this gene causes Fanconi renotubular syndrome, a disease that specifically affects the proximal tubule—and *HNF4A* was recently shown to be required for formation of the proximal tubule in mice (*161*).

Integration of human and mouse developmental atlases

The transition from embryonic to fetal development is of considerable interest, but access to human embryonic tissues is even more limited than access to fetal tissues. To again leverage the mouse, we sought to integrate these human fetal data with a mouse organogenesis cell atlas (MOCA), for which we had previously profiled 2 million cells from undissected embryos spanning E9.5 to E13.5 (*11*). For context, this window corresponds to days 22 to 44 of human development (*162*, *163*), whereas the tissues studied here are estimated to derive from days 72 to 129.

First, we compared the 77 main cell types defined here against the developmental trajectories of organogenesis defined by MOCA by means of a cell type cross-matching method (*11*). Most human cell types strongly matched to a single major mouse trajectory and subtrajectory (fig. S20 and tables S13 and S14). These generally corresponded to expectation, although a few discrepancies facilitated corrections to MOCA (see legends of figs. S20 and S21). Many human cell types and mouse trajectories that lacked strong 1:1 matches [summed nonnegative least squares (NNLS) regression coefficients < 0.6] corresponded to tissues excluded in the other dataset (e.g., mouse placenta and human skin and gonads). Other ambiguities probably follow from the gap between the developmental windows studied (e.g., adrenal cell types), rarity (e.g., bipolar cells), and/or complex developmental relationships (e.g., fetal cell types that derive from multiple embryonic trajectories).

Second, we sought to directly coembed human and mouse cells together. In brief, we sampled 100,000 mouse embryonic cells from MOCA (randomly) and ~65,000 human fetal cells (maximum 1000 cells from each of 77 cell types) and subjected these to integrated analysis (*15*). The distribution of mouse cells in the resulting UMAP visualization was similar to our global analysis of MOCA (Fig. 6, A to C, and figs. S21 to S23) (*11*). Furthermore, despite the species difference, human fetal cells were overwhelmingly distributed in a manner that respected developmental relationships between cell types. For example, human fetal endothelial, hematopoietic, hepatic, epithelial, and mesenchymal cells all mapped to the corresponding mouse embryonic trajectories (Fig. 6B and fig. S21). Within each major trajectory, mouse cells order by successive time point (*11*), whereas human fetal cells appear to project from the last (E13.5) mouse embryonic time point (Fig. 6C). At the subtrajectory level, seniscal mappings include human fetal intestinal epithelial cells emanating from the mouse midgut-hindgut subtrajectory; human fetal parietal and chief cells (stomach) and acinar and ductal cells (pancreas) emanating from the mouse foregut epithelial subtrajectory; human fetal bronchiolar and alveolar epithelial cells emanating from the mouse lung epithelial trajectory; human fetal ureteric bud and metanephric cells emerging separately from the mouse embryonic renal epithelial trajectory; and many others (figs. S21 to S23).

However, there were also a few surprises. For example, although central nervous system (CNS) neurons mapped to the neural tube trajectory and enteric nervous system (ENS) glia and Schwann cells mapped to peripheral nervous system (PNS) glial trajectories, some neural crest derivatives—including ENS neurons, visceral neurons, sympathoblasts, and chromaffin cells—clustered separately from the corresponding mouse embryonic trajectories (figs. S21 to S23), potentially because of excessive differences between the developmental stages or between the species. Human fetal astrocytes clustered with the mouse embryonic neural epithelial trajectory [mouse astrocytes

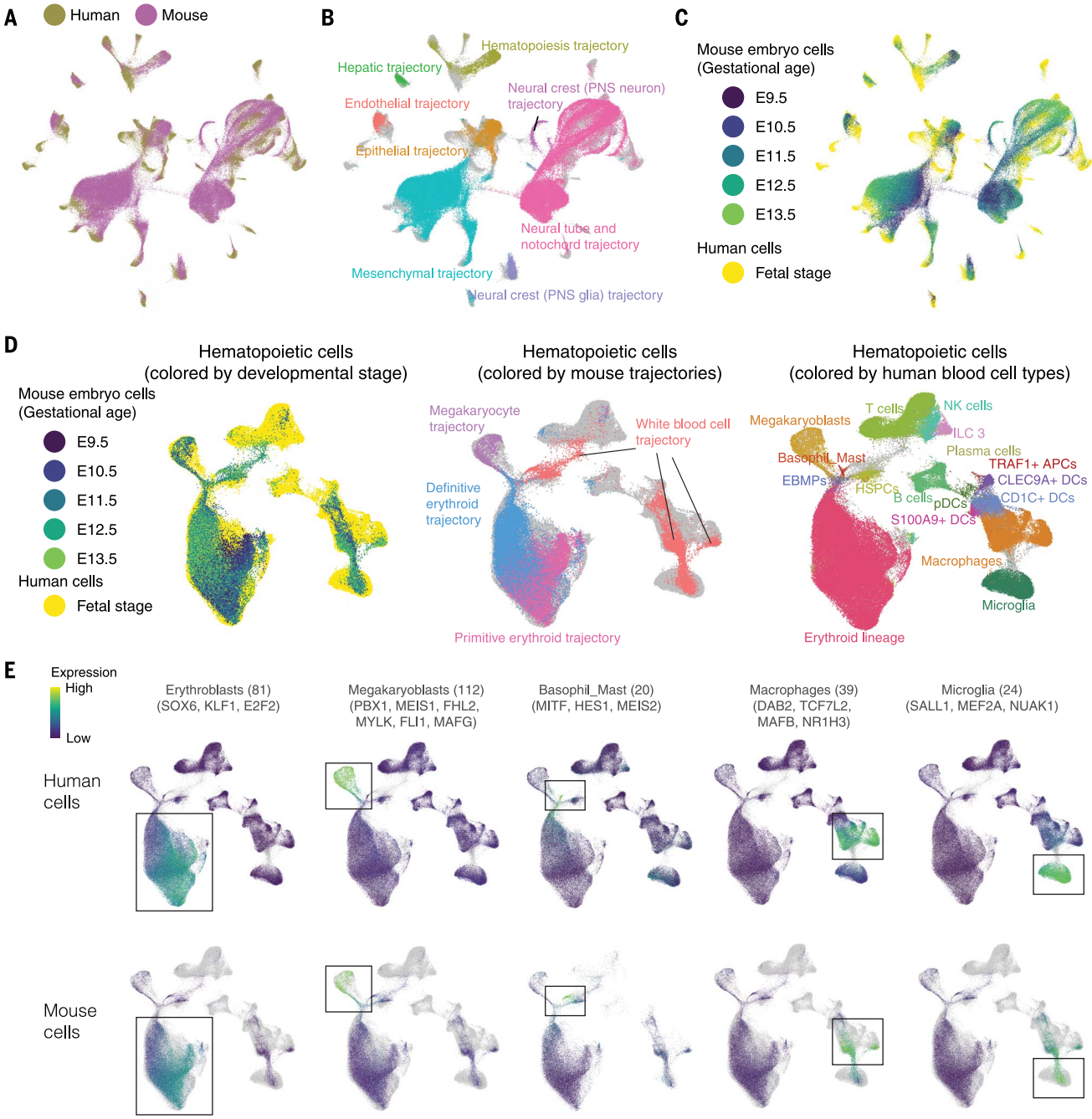


Fig. 6. Integration of human fetal and mouse embryonic cell atlases. (**A** to **C**) After downsampling as described in the text, we applied Seurat (*15*) to jointly analyze human fetal and mouse embryonic cells (*11*). (A) Cells are colored by source species. (B) Mouse cells are colored by the identity of the main mouse embryonic trajectory (*11*). Human cells are colored in gray. (C) Cells are colored by source and development stage. Within each major trajectory and as has been shown previously (*11*), mouse cells order by successive time points, and human fetal cells appear to project from the last (E13.5) mouse embryonic time point. (**D**) We applied Seurat (*15*) to jointly analyze 103,766 human and 40,606 mouse hematopoietic cells. The same UMAP visualization is shown in all panels. (Left) Cells are colored by source and development stage. (Middle) Mouse cells are colored by the identity of mouse subtrajectory (*11*). Human cells are colored in gray. (Right) Human cells are colored according to annotations from Fig. 4B. Mouse cells are colored in gray. (**E**) Plot similar to (D), colored by the normalized expression of human-mouse conserved cell type–specific genes, with their number listed and top TFs named. UMI counts for these genes are scaled for library size, log-transformed, aggregated, and then mapped to *Z* scores.

do not develop until E18.5 (*164*)]. Human fetal oligodendrocytes overlap a rare mouse embryonic subtrajectory (*Pdgfra*+ glia) that, in retrospect, is more likely to correspond to oligodendrocyte precursors (*Olig1*+, *Olig2*+, and *Brinp3*+) (*165*, *166*), which calls into question our previous annotation of a different *Olig1*+ subtrajectory as oligodendrocyte precursors (*11*). These and other unexpected relationships merit further investigation.

To assess relationships between mouse embryonic and human fetal cells in greater detail, we applied the same strategy to extracted cells from the hematopoietic (Fig. 6D and fig. S24), endothelial (fig. S25), and epithelial (fig. S26) trajectories. In these visualizations, we observe examples of the organ-resolved human data deconvoluting the whole-embryo mouse data

into more fine-grained subsets. For example, subsets of the mouse white blood cell embryonic subtrajectory (*11*) map to specific human blood cell types such as HSPCs, microglia, macrophages (liver and spleen), macrophages (other organs), and dendritic cells (DCs) (Fig. 6D). These subsets were further validated by the expression of related blood cell markers (fig. S24C) and annotated on the basis of their human *k*-nearest neighbors ($k = 3$) in the co-embedding (fig. S24D).

Out of 1087 human fetal blood cell type–specific gene markers that are also differentially expressed across mouse blood cell types, 337 genes were differentially expressed (FDR of 0.05) in the same cell type (Fig. 6E and table S15; for comparison, only 12 genes intersected after permutations of labels). In total, 28 of these 337 conserved markers were TFs, 24 of which have been previously reported to be involved in early blood cell differentiation or maintenance for target cell types—e.g., *HLF* as a critical regulator of HSPCs quiescence (*167*), *MITF* as driving mast cell differentiation (*168*), *PAX5* as a master regulator of B cell development (*169*), and *SOX6* as enhancing the differentiation of erythroid progenitors (*170*). However, 4 of the 28 conserved marker TFs have not been previously characterized in the relevant context: *NR1D2* in IL 3 cells, *TCF7L2* in macrophages, *FHL2* in megakaryoblasts, and *NUAK1* in microglia.

In this same analysis, human fetal macrophage and microglia form distinct clusters, but they are connected by a subset of mouse cells from the white blood cell trajectory (Fig. 6D), consistent with previous studies showing that both cell types differentiate from yolk sac progenitors (*171*). To explore this further, we extracted and reanalyzed 4327 mouse embryonic microglia and macrophages by means of unsupervised trajectory analysis (*172*). We observed three smooth cell differentiation trajectories from a common progenitor to microglia in the brain, phagocytic macrophages (*TIMD4*+ and *CD5L*+; mostly in liver, spleen, and adrenal), and perivascular macrophages (*F13A1*+ and *LYVE1*+; widely distributed) (fig. S27A and Fig. 5). The directionality of progression through pseudotime along each macrophage trajectory was consistent with real developmental time (fig. S27B). In total, 1412 genes, including 111 TFs, were differentially expressed in the three macrophage branches (table S16). For example, the microglial trajectory showed elevated expression of *BACH2* and *RUNX3* as well as known microglial regulators *SALL1* (*173*) and *MEF2A* (*173*, *174*), perivascular macrophages of *DAB2*, and *TCF7L2*, and phagocytic macrophages of *MAFB* and *NR1H3* (fig. S27C). Overall, these analyses illustrate how fetal annotations can be used to identify and characterize progenitors of specific lineages at developmental time points where they may be difficult to resolve on their own, even across species.

Discussion

Two centuries after the formulation of the cell theory—the assertion that all living things consist of cells and that the cell is the most basic unit of life (*175*)—we are on the cusp of cataloging and characterizing all cell types that constitute a human body, both in health and disease. To this end, the field of single-cell biology is progressing at an astonishing rate, propelled by a synergy between new technologies and new computational methods to make sense of the data produced by those technologies. In the past few years alone, this synergy has enabled compelling and informative single-cell atlases of many human organs as well as of entire model organisms (*11*, *51*, *69*, *108*, *152*, *176*–*182*).

Human development is a remarkable process that begins with a fertilized zygote and proceeds through a germinal stage followed by embryogenesis. By the end of the 10th week, the embryo has acquired its basic form and is termed a fetus. For the following 30 weeks, all organs continue to grow and mature, with diverse terminally differentiated cell types arising from their progenitors. Although the germinal and embryogenesis stages have been intensively profiled with single-cell methods in humans and mice (*11*, *180*, *181*), it has been more challenging to profile the fetal stage. Although several single-cell studies of human fetal development have recently appeared (*152*, *182*–*184*), these are restricted to individual organs or cell lineages and do not obtain a comprehensive view.

In this study, together with (*12*), we set out to generate single-cell atlases of gene expression and chromatin accessibility using diverse tissues obtained during human fetal development. From 15 distinct organs, we successfully profiled gene expression in ~4 million single cells and chromatin accessibility in ~800,000 single cells. Limitations of these datasets include nonuniform sampling (i.e., more cells profiled in some organs than others), missing tissues (most notably, bone marrow, skin, bone, and gonads), relatively low sequencing depth, and the sparsity of single-cell molecular profiles. Nonetheless, we identified hundreds of cell types and subtypes that are supported by a framework for quantifying specificity as well as by matching nearly all of them to cell types or subtypes from published mouse atlases.

In contrast with organ-specific studies, the diversity of tissues profiled here enabled cross-tissue comparisons of broadly distributed cell types. We emphasize that our process for annotating cell types benefited tremendously from the myriad single-cell atlases of specific human organs or other mammals that have been generated to date (*8*, *9*, *11*, *16*, *28*, *50*, *108*, *139*). Of course, decisions in the annotation process can be subjective (e.g., over- versus under-clustering), and both cell type and subtype annotations made here should be considered preliminary and subject to revision.

The apparent hematopoiesis that we observe in the fetal adrenal gland is consistent with the fact that the adrenal gland, along with many other organs (e.g., spleen, liver, and lymph nodes), can serve as a site of extramedullary hematopoiesis in adults with pathologic conditions that lead to an increased demand for blood cell production, particularly hemoglobinopathies (*185*, *186*). Although occasional islands of extramedullary hematopoiesis have been seen in the adrenal glands of human embryos (*126*, *127*), our findings in both the human and mouse provide quantitative evidence that the adrenal gland serves as a normal, albeit minor, site of erythropoiesis during a developmental window that overlaps with the transition of hematopoiesis from the liver to the marrow.

The ease with which we were able to integrate single-cell profiles from mouse organogenesis and human fetal development is notable, particularly given that these represent different stages of mammalian development, not to mention our separation from mice by >100 million years of evolution. The relatively straightforward alignment of the datasets highlights the extent of evolutionary constraint on the molecular programs of individual cell types, and it furthermore lends support to long-standing use of the mouse as a powerful model system for studying human development.

Looking forward, we envision that the somewhat narrow window of midgestational human development studied here will be complemented by additional atlases of earlier and later time points (e.g., embryonic and adult) as well as by similarly comprehensive profiling and integration of data from model organisms. The continued development and application of methods for ascertaining gene expression and chromatin accessibility—in concert with spatial, epigenetic, proteomic, lineage history, and other information—will be necessary to obtain a comprehensive view of temporal unfolding of human cell type diversity that begins at the single-cell zygote.

To date, investigations of human development have largely been indirect, with key molecular factors nominated by human genetics and then investigated in model organisms and/or in vitro systems. Knowledge of the in vivo landscape of gene expression and regulation has been limited. In filling part of this gap, we hope that this atlas will enable a better understanding of the molecular and cellular basis of both rare and common disorders of human development, while also informing the path to successful therapies.

Materials and methods

A more detailed version of the materials and methods is provided in the supplementary materials.

sci-RNA-seq3

A more detailed version of the full sci-RNA-seq3 workflow is available on protocols.io (*187*) and in the supplementary materials.

Preparation of nuclei

Human fetal tissues (89 to 125 days estimated postconceptual age) were obtained by the University of Washington Birth Defects Research Laboratory (BDRL) under a protocol approved by the University of Washington Institutional Review Board. Tissues of interest were isolated and rinsed in 1X HBSS. Dried tissue was snap frozen in liquid nitrogen, manually pulverized on dry ice with a chilled hammer, aliquoted, and stored at −80°C until further processing. A subset of these aliquots were used for sci-RNA-seq3, and others for sci-ATAC-seq3, as described in the companion paper. For RNA-seq, nuclei from tissues and control cell lines were lysed in the cell lysis buffer and fixed with ice-cold 4% paraformaldehyde (EMS, 15-4-100) on the basis of the published sci-RNA-seq3 protocol (*11*). For human cell extraction in renal and digestive organs (kidney, pancreas, intestine, and stomach) and paraformaldehyde fixation, we followed the procedure described in (*13*).

Immunohistochemistry

Fetal tissues were fixed in formalin and embedded in paraffin. Sections of 4- to 5-μm thickness were cut and placed on Superfrost Plus slides (12-550-17, FisherBrand). For immunohistochemistry, sections were subjected to heat-mediated antigen retrieval (pH 6.0) followed by blocking with normal serum. Primary antibodies were incubated overnight at 4°C. The primary antibody we used: GYPA (R&D, MAB1228, 1:250), CD34 (R&D, AF7227, 1:250), CD34 (Novus, NBP2-32933, 1:250), ANXA1 (R&D, AF3770, 1:500), TNFRS10C (R&D, MAB6301, 1:500), AFP (Novus, NBP1-76275, 1:400), ALB (R&D, MAB1455, 1:10K), AHSG (R&D, AF1184, 1:400), and APOA1 (R&D, MAB36641, 1:250). Species and subtype-appropriate fluorescent dye-labeled secondary antibodies were used (Alexa Fluor 488 and 594, 1:400, Jackson ImmunoResearch Lab) or biotinylated secondary antibody were used followed by ABC Elite Systems (PK-6100, Vector Lab) for 3,3′-diaminobenzidine (DAB) chromogen staining.

sci-RNA-seq3 library construction and sequencing

The paraformaldehyde fixed nuclei were processed similarly to the published sci-RNA-seq3 protocol (*11*). For paraformaldehyde fixed cells, frozen fixed cells were thawed on 37°C water bath, spun down at 500 × *g* for 5 min, and incubated with 500 μl PBSI [1 x phosphate-buffered saline (PBS), pH 7.4, 1% bovine serum albumin (BSA), 1% SuperRnaseIn] including 0.2% Triton X-100 for 3 min on ice. Cells were pelleted and resuspended in 500 μl nuclease-free water including 1% SuperRnaseIn. 3 ml 0.1N HCl were added into the cells for 5min incubation on ice (*17*). 3.5 ml Tris-HCl (pH 8.0) and 35 μl 10% Triton X-100 were added into cells to neutralize HCl. Cells were pelleted and washed with 1 ml PBSR. Cells were pelleted and resuspended in 100 μl PBSI. The following steps were similar with the sci-RNA-seq3 protocol (with paraformaldehyde fixed nuclei) with slight modifications: (i) We distributed 20,000 fixed cells (instead of 80,000 nuclei) per well for reverse transcription (RT). (ii) We replaced all nuclei wash buffer in following steps with PBSI. (iii) All nuclei dilution buffer were replaced with PBS + 1% BSA.

Processing of sequencing reads

Read alignment and gene count matrix generation for the scRNA-seq was performed using the pipeline that we developed for sci-RNA-seq3 (*11*) with minor modifications: Duplicates were removed using the UMI sequence (ED < 2, including insertions and deletions), RT index, hairpin ligation adaptor index, and read 2 end-coordinate.

After the single-cell gene count matrix was generated, cells with <250 UMIs were filtered out. Each cell was assigned to its original human fetal sample on the basis of the RT barcode. Reads mapping to each fetus individual were aggregated to generate pseudobulk RNA-seq datasets. For sex assignments, we counted reads mapping to female-specific noncoding RNA (*TSIX* and *XIST*) or chrY genes (except genes *TBL1Y*, *RP11-424G14.1*, *NLGN4Y*, *AC010084.1*, *CD24P4*, *PCDH11Y*, and *TTTY14*, which are detected in both males and females). Fetuses were readily separated into females (more reads mapping to *TSIX* and *XIST* than chrY genes) and males (more reads mapping to chrY genes than *TSIX* and *XIST*).

Clustering analysis of pseudobulk transcriptomes was done with Monocle 3/alpha (*11*). Briefly, an aggregated gene expression matrix was constructed as described above for human fetal organs from each individual. Samples with >5000 total UMIs were selected. The dimensionality of the data was reduced by principal components analysis (PCA) (10 components), first on the top 500 most highly dispersed genes and then with UMAP (max_components = 2, n_neighbors = 10, min_dist = 0.5, metric = 'cosine').

Cell filtering, clustering and marker gene identification

For the detection of potential doublet cells and doublet-derived subclusters from each organ, we used an iterative clustering strategy as shown before (*11*). For data visualization, cells labeled as doublets [by scrublet/v0.1 pipeline (*188*)] or from doublet-derived subclusters were filtered out. For each cell, we only retain protein-coding genes, lincRNA genes and pseudogenes. Genes expressed in <10 cells and cells expressing <100 genes were further filtered out. The downstream dimension reduction and clustering analysis were done by Monocle 3/alpha with similar settings (*11*). Clusters were assigned to known cell types on the basis of cell type–specific markers (table S3). We found the above Scrublet and iterative clustering-based approach is limited in marking cell doublets between abundant cell clusters and rare cell clusters (e.g., <1% of total cell population). To further remove such doublet cells, we took the cell clusters identified by Monocle 3 and first computed differentially expressed genes across cell clusters (within-organ) with the differentialGeneTest() function of Monocle 3. We then selected a gene set combining the top ten gene markers for each cell cluster (ordered by *q* value and fold expression difference between first and second ranked cell cluster). Cells from each main cell cluster were selected for dimension reduction by PCA (10 components) first on the selected gene set of top cluster specific gene markers, and then by UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'), followed by clustering identification using the density peak clustering algorithm implemented in Monocle 3 (rho_thresh = 5, delta_thresh = 0.2 for most clustering analysis). Subclusters showing low expression of target cell cluster specific markers and enriched expression of nontarget cell cluster specific markers were annotated as doublets derived subclusters and filtered out in visualization and downstream analysis. Differentially expressed genes across cell types (within-organ) were recomputed with the differentialGeneTest() function of Monocle 3 after removing all doublets or cells from doublet-derived subclusters.

Adjudication of the 15 initially unannotated cell types

As noted in the main text, our first round of annotation was performed on a tissue-by-tissue basis by comparing observed cell types with those expected from prior knowledge of the same tissue. In general, we recovered all or nearly all main cell types identified by previous atlasing efforts directed at the same organs, despite differences with respect to species, stage of development and/or technology. Additionally, we identified 15 cell types that we did not at least initially expect to observe in a given tissue. We labeled these on the basis of the top enriched differentially expressed gene markers within that tissue, e.g., CSH1_CSH2 positive cells. After the initial round of annotation, we reexamined these 15 cell types on the basis of their distribution

in the global UMAP, whether they matched annotated cell types in mouse atlases, their distribution across tissues derived from different individuals, and their potential for maternal origin. Our updated interpretations are summarized in the supplementary materials.

Clustering analysis of cells across organs

For clustering analysis of 77 main cell types across 15 organs, we sampled 5000 cells from each cell type (or all cells for cell types with <5000 cells in a given organ). The dimensionality of the data was reduced first by PCA (50 components) on the gene set combining top cell type–specific gene markers identified above (table S5, q value = 0) and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). Differentially expressed genes across cell types were identified with the differentialGeneTest() function of Monocle 3. For annotating cell type–specific gene features, we intersected the cell type–specific genes identified above with the predicted secreted and membrane protein coding gene sets from the Human Protein Atlas (*189*), as well as the TF set annotated in the "motifAnnotations_hgnc" data from package RcisTarget/v1.2.1 (*190*).

For clustering analysis of blood cell across 15 organs, we extracted all blood cells corresponding to annotated clusters of myeloid cells, lymphoid cells, thymocytes, megakaryocytes, microglia, antigen presenting cells, erythroblasts, and HSPCs. The dimensionality of the data was reduced first by PCA (40 components) on the expression of a gene set combining the top 3000 blood cell type–specific gene markers (table S5, only genes specifically expressed in at least one blood cell type were selected (q < 0.05, fold expression difference between first and second ranked cell cluster > 2) and ordered by median q value across organs) and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). Cell clusters were identified using the Louvain algorithm implemented in Monocle 3 (louvain_res = 1×10^{-4}). Clusters were assigned to known cell types on the basis of cell type–specific markers. We then coembedded the human fetal blood cells and a scRNA-seq atlas of blood cells from the fetal liver (*108*), using the Seurat v3 integration method (FindAnchors and IntegrateData) (*15*) with a chosen dimensionality of 30 on the top 3000 highly variable genes with shared gene names in both datasets.

We then applied a similar analysis strategy as above for clustering analysis of endothelial or epithelial cells across organs. For endothelial cells, we first extracted cells corresponding to annotated clusters of vascular endothelial cells, lymphatic endothelial cells and endocardial cells across organs. The dimensionality of the data was reduced first by PCA (30 components) on the gene set combining top 1000 endothelial cell type–specific gene markers identified above (table S5, only genes specifically expressed in at least one endothelial cell type were selected (q < 0.05, fold expression difference between first and second ranked cell cluster > 2) and ordered by median q value across organs) and then with UMAP with the same parameters used for blood cells. Cell clusters were identified using the Louvain algorithm implemented in Monocle 3 (louvain_res = 1×10^{-4}), and then annotated on the basis of the tissue origin of endothelial cells. For epithelial cells, we first extracted cells from the epithelial cell cluster in fig. S4B, followed by dimension reduction first by PCA (50 components) first on the top 5000 most highly dispersed genes and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). For validating the tissue specific endothelial cells, we then coembedded the human fetal endothelial cells and a scRNA-seq atlas of endothelial cells from mouse adult tissues (*139*), using the Seurat v3 integration method (FindAnchors and IntegrateData) (*15*) with a chosen dimensionality of 30 on the top 3000 highly variable genes with shared gene names in both datasets.

Intradataset cross-validation analysis

For cells from each organ, we randomly sampled up to 2000 cells from each main cell type. We then followed the same process (*101*). Briefly, we combined all sampled cells from each organ and evaluated cell type specificity by applying a fivefold cross-validation to the dataset, with an SVM classifier (with linear kernel). Whole transcriptome was used in cell type prediction. We then computed the cross-validation F1 value as cell type specificity score. As control, we randomly permuted the cell type labels, followed by the same analysis pipeline. For cell type specificity analysis across all organs, we applied the same analysis strategy to the full dataset after sampling up to 2000 cells of each main cell type.

Subclustering analysis

For each main cell type (with >1000 cells) in each organ, we applied Harmony/v1.0 for batch correction and dimension reduction (*102*). Briefly, the dimensionality of the data was reduced by PCA (30 components, or 10 components for cell types with <5000 cells) first on the top 3000 (or 1000 for cell types with <5000 cells) most highly variable genes, followed by batch correction on sample ID. Cell clusters were identified using the Louvain algorithm implemented in Seurat/v3.1.4 (*15*) (resolution = 0.5). We then applied the intradataset cross-validation approach to evaluate the specificity of subclusters within each main cell type. For every subcluster pair, A and B, we computed the number of A cells mislabeled as B cells in cross-validation analysis with the true dataset (mislabeled cell number: n) or the permuted dataset (mislabeled cell number: m). A large n value suggests the two subclusters are not well separated by the full transcriptome. We thus iteratively merged similar subcluster pairs ($n > m$), and identified a total of 657 subtypes across 15 organs. The intradataset cross-validation approach was applied to evaluating subtype specificity within each main cell type in each organ. To annotate the identity of subtypes, we applied the same cell type correlation analysis strategy described in (*11*) to compare cell subtypes from this study with cell types of the same organ from the Microwell-seq based Mouse Cell Atlas (MCA) (*16*). A similar comparison was performed for all subtypes from the brain against cell types annotated in a recent mouse brain atlas (MBCA) (*50*).

Validating erythropoiesis in the adrenal tissues from newborn mice

Adrenals and kidneys were harvested from CD1 Swiss albino mice (Charles River) on the day of birth (P0), and bone marrow cells were flushed from the femurs of the dams. Solid tissues were dissociated using collagenase and stained for imaging flow cytometry using the markers Ter119 (AF488), CD117 (PE-CF594), CD71 (PE), CD45 (EF450), and DRAQ5. Gating of maturing erythroblast populations was performed using published methods (*128*) and analyzed with IDEAS (Luminex) software.

Comparison of human and mouse developmental atlases

We first applied a slightly modified version of the strategy described in (*11*) to identify correlated cell types between this human fetal cell atlas and the mouse organogenesis cell atlas (MOCA) (*11*). As a different approach, we coembedded the human fetal cell atlas and the mouse organogenesis cell atlas (MOCA) (*11*) using the Seurat v3 integration method (FindAnchors and IntegrateData) (*15*) with a chosen dimensionality of 30 on the top 3000 highly variable genes with shared gene names in both human and mouse. We first integrated 65,000 human fetal cells (up to 1000 cells randomly sampled from each of 77 cell types) and 100,000 mouse embryonic cells (randomly sampled from MOCA) with default parameters. We then applied the same integrative analysis strategy to extracted human and mouse cells from the hematopoietic, endothelial, and epithelial trajectories.

For the coembedded human and mouse hematopoietic cells, we annotated each mouse cell on the basis of its k-nearest neighbors of human cells. We chose a small k value (k = 3) such that rare cell types were also annotated. Differentially expressed genes across mouse hematopoietic cells were computed with the

differentialGeneTest() function of Monocle 3/alpha.

Pseudotemporal ordering of mouse macrophage/microglia cells was done with Monocle 3/alpha with the reduction method of "DDRTree." Briefly, the top 3 principal components on the top 500 highly variable genes were used to construct the DDRTree pseudotime trajectory with UMI number per cell as a covariate [param.gamma = 120, norm_method = "log," residualModelFormulaStr = "~ sm.ns(Total_mRNAs, df = 3)"]. The cells are separated into three branch trajectories in the DDRTree space. Differentially expressed genes across the three branches were computed with the differentialGeneTest() function of Monocle 3/alpha. We then clustered cells with k means clustering (k = 10) and computed the average development time for each cluster. The progenitor cell group was annotated on the basis of the lowest average development time and appeared at the center of the three branches. Each cell was assigned a pseudotime value on the basis of its distance from the progenitor cells.

Using the Garnett models trained on this human cell atlas for cell type classification

The R package Garnett for Monocle 3 (version 0.2.9) was used to generate cell type classifiers for each of the 15 tissues. Marker genes for each cell type were assembled from literature searches, and models were trained using train_cell_classifier using default parameters and the gene database org.Hs.eg.db (version 3.10.0). Models were trained on the entirety of each tissue dataset with the exception of cerebrum, where 100,000 cells were randomly sampled for training for computational efficiency. The automatic and manual annotation processes were conducted independently. To compare cell type assignments with those obtained through manual annotation (i.e., the 77 main cell types), we applied the function classify_cells using the trained models with the following nondefault parameters: cluster_extend = TRUE, cluster_extend_max_frac_incorrect = 0.25, cluster_extend_max_frac_unknown = 0.95. Garnett cell type assignments that matched the cell type assignment from manual annotation were considered correct with the following exceptions: Garnett classification of "Chromaffin cells" was considered correct when manual annotation was "Sympathoblasts;" Garnett classification of "B cells" or "T cells" was considered correct when manual annotation was "Lymphoid cells;" Garnett classification of "Cap mesenchyme cells," "Collecting duct cells," "Distal tubule cells," "Loop of Henle cells," "Proximal tubule cells," and "Podocytes" were considered correct when manual annotation was "Metanephric cells;" Garnett classification of "Ureter cells" and "Collecting duct cells" were considered correct when manual annotation was "Ureteric bud cells;" Garnett classification of "Pancreatic Alpha cells," "Pancreatic Beta cells," and "Pancreatic Delta cells" was considered correct when manual annotation was "Islet endocrine cells;" Garnett classification of "D cells" was considered correct with manual annotation of "Neuroendocrine cells."

To test the applicability of Garnett trained models to future data, we applied the pancreas model to human adult pancreas scRNA-seq data from reference (*100*). The model was applied using the function classify_cells with the same parameters as above. When comparing cell type assignments with those provided by the authors, we considered the following cell types to be equivalent: acinar, Acinar cells; ductal, Ductal cells; endothelial, Endothelial cells; mast, Myeloid cells; macrophage, Myeloid cells; schwann, Glia; alpha, Pancreatic Alpha cells; beta, Pancreatic Beta cells; delta, Pancreatic Delta cells; activated_stellate, Pancreatic stellate cells; quiescent_stellate, Pancreatic stellate cells; and t_cell, T cells.

REFERENCES AND NOTES

1. B. S. de Bakker *et al*., An interactive three-dimensional digital atlas and quantitative database of human development. *Science* **354**, aag0053 (2016). doi: 10.1126/science.aag0053; pmid: 27884980
2. S. Yamada *et al*., Developmental atlas of the early first trimester human embryo. *Dev. Dyn.* **239**, 1585–1595 (2010). doi: 10.1002/dvdy.22316; pmid: 20503356
3. J. E. Jirasek, *An Atlas of the Human Embryo and Fetus: A Photographic Review of Human Prenatal Development* (CRC Press, 2000).
4. V. A. McKusick, Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.* **80**, 588–604 (2007). doi: 10.1086/514346; pmid: 17357067
5. H. C. Mefford, M. L. Batshaw, E. P. Hoffman, Genomics, intellectual disability, and autism. *N. Engl. J. Med.* **366**, 733–743 (2012). doi: 10.1056/NEJMra1114194; pmid: 22356326
6. S. E. McCandless, J. W. Brunger, S. B. Cassidy, The burden of genetic disease on inpatient care in a children's hospital. *Am. J. Hum. Genet.* **74**, 121–127 (2004). doi: 10.1086/381053; pmid: 14681831
7. M. H. Wojcik *et al*., Genetic disorders and mortality in infancy and early childhood: Delayed diagnoses and missed opportunities. *Genet. Med.* **20**, 1396–1404 (2018). doi: 10.1038/gim.2018.17; pmid: 29790870
8. S. Darmanis *et al*., A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 7285–7290 (2015). doi: 10.1073/pnas.1507125112; pmid: 26060301
9. E. Madissoon *et al*., scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol.* **21**, 1 (2019). doi: 10.1186/s13059-019-1906-x; pmid: 31892341
10. S. Behjati, S. Lindsay, S. A. Teichmann, M. Haniffa, Mapping human development at single-cell resolution. *Development* **145**, dev152561 (2018). doi: 10.1242/dev.152561; pmid: 29439135
11. J. Cao *et al*., The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019). doi: 10.1038/s41586-019-0969-x; pmid: 30787437
12. S. Domcke *et al*., A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020). doi: 10.1126/science.aba7612
13. E. R. Thomsen *et al*., Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat. Methods* **13**, 87–93 (2016). doi: 10.1038/nmeth.3629; pmid: 26524239
14. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018). doi: 10.21105/joss.00861
15. T. Stuart *et al*., Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019). doi: 10.1016/j.cell.2019.05.031; pmid: 31178118
16. X. Han *et al*., Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091–1107.e17 (2018). doi: 10.1016/j.cell.2018.05.012; pmid: 29775597
17. A. B. Rosenberg *et al*., Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018). doi: 10.1126/science.aam8999; pmid: 29545511
18. J. Cao *et al*., Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017). doi: 10.1126/science.aam8940; pmid: 28818938
19. M. Plass *et al*., Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018). doi: 10.1126/science.aaq1723; pmid: 29674432
20. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281–291.e9 (2019). doi: 10.1016/j.cels.2018.11.005; pmid: 30954476
21. G. Stelzer *et al*., The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinformatics* **54**, 1.30.1–1.30.33 (2016). doi: 10.1002/cpbi.5
22. G. Teitelman, H. Baker, T. H. Joh, D. J. Reis, in *Catecholamines: Basic and Clinical Frontiers*, E. Usdin, I. J. Kopin, J. Barchas, Eds. (Pergamon, 1979), pp. 794–796.
23. Y. Lu, W. Yi, Q. Wu, S. Zhong, Z. Zuo, F. Zhao, M. Zhang, N. Tsai, Y. Zhuo, S. He, J. Zhang, X. Duan, X. Wang, T. Xue, Single-cell RNA-seq analysis maps the development of human fetal retina. bioRxiv 423830 [Preprint]. 23 September 2018. doi: 10.1101/423830
24. The UniProt Consortium, UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017). doi: 10.1093/nar/gkw1099; pmid: 27899622
25. Gene_Cryab (2020); www.ncbi.nlm.nih.gov/gene/12955.
26. CRBB3_HUMAN, UniProt accession no. P26998; www.uniprot.org/uniprot/P26998.
27. K1C12_HUMAN, UniProt accession no. Q99456; www.uniprot.org/uniprot/Q99456.
28. Tabula Muris Consortium, Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018). doi: 10.1038/s41586-018-0590-4; pmid: 30283141
29. B. R. Kim *et al*., SOX2 and PI3K Cooperate to Induce and Stabilize a Squamous-Committed Stem Cell Injury State during Lung Squamous Cell Carcinoma Pathogenesis. *PLOS Biol.* **14**, e1002581 (2016). doi: 10.1371/journal.pbio.1002581; pmid: 27880766
30. S. Dauger *et al*., Phox2b controls the development of peripheral chemoreceptors and afferent visceral pathways. *Development* **130**, 6635–6642 (2003). doi: 10.1242/dev.00866; pmid: 14627719
31. D. G. Jackson, R. Prevo, S. Clasper, S. Banerji, LYVE-1, the lymphatic system and tumor lymphangiogenesis. *Trends Immunol.* **22**, 317–321 (2001). doi: 10.1016/S1471-4906(01)01936-6; pmid: 11377291
32. Y. Rinkevich *et al*., Identification and prospective isolation of a mesothelial precursor lineage giving rise to smooth muscle cells and fibroblasts for mammalian internal organs, and their vasculature. *Nat. Cell Biol.* **14**, 1251–1260 (2012). doi: 10.1038/ncb2610; pmid: 23143399
33. Y. Cui *et al*., Single-Cell Transcriptome Analysis Maps the Developmental Track of the Human Heart. *Cell Rep.* **26**, 1934–1950.e5 (2019). doi: 10.1016/j.celrep.2019.01.079; pmid: 30759401
34. M. E. Shy *et al*., Schwann cell expression of PLP1 but not DM20 is necessary to prevent neuropathy. *Ann. Neurol.* **53**, 354–365 (2003). doi: 10.1002/ana.10466; pmid: 12601703
35. H. Suryawanshi *et al*., A single-cell survey of the human first-trimester placenta and decidua. *Sci. Adv.* **4**, eaau4788 (2018). doi: 10.1126/sciadv.aau4788; pmid: 30402542
36. F. E. Thorngate, P. A. Strockbine, S. K. Erickson, D. L. Williams, Altered adrenal gland cholesterol metabolism in the apoE-deficient mouse. *J. Lipid Res.* **43**, 1920–1926 (2002). doi: 10.1194/jlr.M200205-JLR200; pmid: 12401891
37. A. Furlan *et al*., Multipotent peripheral glial cells generate neuroendocrine cells of the adrenal medulla. *Science* **357**, eaal3753 (2017). doi: 10.1126/science.aal3753; pmid: 28684471
38. M. E. Fomin, A. I. Beyer, M. O. Muench, Human fetal liver cultures support multiple cell lineages that can engraft immunodeficient mice. *Open Biol.* **7**, 170108 (2017). doi: 10.1098/rsob.170108; pmid: 29237808
39. S. Matsuoka *et al*., CD34 expression on long-term repopulating hematopoietic stem cells changes during developmental stages. *Blood* **97**, 419–425 (2001). doi: 10.1182/blood.v97.2.419; pmid: 11154218

40. J. D. Miller, S. M. Lankford, K. B. Adler, A. R. Brody, Mesenchymal stem cells require MARCKS protein for directed chemotaxis in vitro. *Am. J. Respir. Cell Mol. Biol.* **43**, 253–258 (2010). doi: 10.1165/rcmb.2010-0015RC; pmid: 20224071
41. F. Dijk, A. A. B. Bergen, W. Kamphuis, GAP-43 expression is upregulated in retinal ganglion cells after ischemia/reperfusion-induced damage. *Exp. Eye Res.* **84**, 858–867 (2007). doi: 10.1016/j.exer.2007.01.006; pmid: 17343850
42. R. P. de Melo Guimarães *et al*., Evidence of Müller Glia Conversion Into Retina Ganglion Cells Using Neurogenin2. *Front. Cell. Neurosci.* **12**, 410 (2018). doi: 10.3389/fncel.2018.00410; pmid: 30483060
43. R. A. Poché *et al*., Lim1 is essential for the correct laminar positioning of retinal horizontal cells. *J. Neurosci.* **27**, 14099–14107 (2007). doi: 10.1523/JNEUROSCI.4046-07.2007; pmid: 18094249
44. P. A. Ruzycki, X. Zhang, S. Chen, CRX directs photoreceptor differentiation by accelerating chromatin remodeling at specific target sites. *Epigenetics Chromatin* **11**, 42 (2018). doi: 10.1186/s13072-018-0212-2; pmid: 30068366
45. Z. Shi *et al*., Vsx1 regulates terminal differentiation of type 7 ON bipolar cells. *J. Neurosci.* **31**, 13118–13127 (2011). doi: 10.1523/JNEUROSCI.2331-11.2011; pmid: 21917795
46. M. V. Sofroniew, H. V. Vinters, Astrocytes: Biology and pathology. *Acta Neuropathol.* **119**, 7–35 (2010). doi: 10.1007/s00401-009-0619-8; pmid: 20012068
47. F. Beleggia *et al*., CRIM1 haploinsufficiency causes defects in eye development in human and mouse. *Hum. Mol. Genet.* **24**, 2267–2273 (2015). doi: 10.1093/hmg/ddu744; pmid: 25561690
48. C. J. Jaworski *et al*., Expression analysis of human pterygium shows a predominance of conjunctival and limbal markers and genes associated with cell migration. *Mol. Vis.* **15**, 2421–2434 (2009). pmid: 19956562
49. S. Rétaux, M. Rogard, I. Bach, V. Failli, M. J. Besson, *Lhx9*: A novel LIM-homeodomain gene expressed in the developing forebrain. *J. Neurosci.* **19**, 783–793 (1999). doi: 10.1523/JNEUROSCI.19-02-00783.1999; pmid: 9880598
50. A. Zeisel *et al*., Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999–1014.e22 (2018). doi: 10.1016/j.cell.2018.06.021; pmid: 30096314
51. R. Menon *et al*., Single-cell analysis of progenitor cell dynamics and lineage specification in the human fetal kidney. *Development* **145**, dev164038 (2018). doi: 10.1242/dev.164038; pmid: 30166318
52. M. Ben Khelifa *et al*., Mutations in *DNAH1*, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. *Am. J. Hum. Genet.* **94**, 95–104 (2014). doi: 10.1016/j.ajhg.2013.11.017; pmid: 24360805
53. H. He *et al*., Krüppel-like Factor 4 Promotes Esophageal Squamous Cell Carcinoma Differentiation by Up-regulating Keratin 13 Expression. *J. Biol. Chem.* **290**, 13567–13577 (2015). doi: 10.1074/jbc.M114.629717; pmid: 25851906
54. Y. Pan, W.-D. Wang, T. Yago, Transcriptional regulation of podoplanin expression by Prox1 in lymphatic endothelial cells. *Microvasc. Res.* **94**, 96–102 (2014). doi: 10.1016/j.mvr.2014.05.006; pmid: 24944097
55. U. Hopfer *et al*., Targeted disruption of *Col8a1* and *Col8a2* genes in mice leads to anterior segment abnormalities in the eye. *FASEB J.* **19**, 1232–1244 (2005). doi: 10.1096/fj.04-3019com; pmid: 16051690
56. R. Vento-Tormo *et al*., Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018). doi: 10.1038/s41586-018-0698-6; pmid: 30429548
57. J. A. Gomez-Sanchez *et al*., Schwann cell autophagy, myelinophagy, initiates myelin clearance from injured nerves. *J. Cell Biol.* **210**, 153–168 (2015). doi: 10.1083/jcb.201503019; pmid: 26150392
58. M. Hoekstra, M. Van Eck, HDL is redundant for adrenal steroidogenesis in LDLR knockout mice with a human-like lipoprotein profile. *J. Lipid Res.* **57**, 631–637 (2016). doi: 10.1194/jlr.M066019; pmid: 26891738
59. K. Huber, N. Karch, U. Ernsberger, C. Goridis, K. Unsicker, The role of Phox2B in chromaffin cell development. *Dev. Biol.* **279**, 501–508 (2005). doi: 10.1016/j.ydbio.2005.01.007; pmid: 15733675
60. X. Pei, Who is hematopoietic stem cell: CD34+ or CD34-? *Int. J. Hematol.* **70**, 213–215 (1999). pmid: 10643145
61. J. R. Lavoie *et al*., Brief Report: Elastin Microfibril Interface 1 and Integrin-Linked Protein Kinase Are Novel Markers of Islet Regenerative Function in Human Multipotent Mesenchymal Stromal Cells. *Stem Cells* **34**, 2249–2255 (2016). doi: 10.1002/stem.2385; pmid: 27090767
62. D. S. Wall *et al*., Progenitor cell proliferation in the retina is dependent on Notch-independent Sonic hedgehog/Hes1 activity. *J. Cell Biol.* **184**, 101–112 (2009). doi: 10.1083/jcb.200805155; pmid: 19124651
63. F. Wu *et al*., Two transcription factors, Pou4f2 and Isl1, are sufficient to specify the retinal ganglion cell fate. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1559–E1568 (2015). doi: 10.1073/pnas.1421535112; pmid: 25775587
64. R. Balasubramanian, A. Bui, X. Dong, L. Gan, Lhx9 Is Required for the Development of Retinal Nitric Oxide-Synthesizing Amacrine Cell Subtype. *Mol. Neurobiol.* **55**, 2922–2933 (2018). doi: 10.1007/s12035-017-0554-y; pmid: 28456934
65. F. Wu *et al*., Onecut1 is essential for horizontal cell genesis and retinal integrity. *J. Neurosci.* **33**, 13053–13065 (2013). doi: 10.1523/JNEUROSCI.0116-13.2013; pmid: 23926259
66. M. Vitorino *et al*., Vsx2 in the zebrafish retina: Restricted lineages through derepression. *Neural Dev.* **4**, 14 (2009). doi: 10.1186/1749-8104-4-14; pmid: 19344499
67. S. Kolandaivelu, V. Ramamurthy, AIPL1 protein and its indispensable role in cone photoreceptor function and survival. *Adv. Exp. Med. Biol.* **801**, 43–48 (2014). doi: 10.1007/978-1-4614-3209-8_6; pmid: 24664679
68. I. R. Holtman *et al*., Induction of a common microglia gene expression signature by aging and neurodegenerative conditions: A co-expression meta-analysis. *Acta Neuropathol. Commun.* **3**, 31 (2015). doi: 10.1186/s40478-015-0203-5; pmid: 26001565
69. S. Zhong *et al*., A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018). doi: 10.1038/nature25980; pmid: 29539641
70. M. Lee *et al*., Tcf7l2 plays crucial roles in forebrain development through regulation of thalamic and habenular neuron identity and connectivity. *Dev. Biol.* **424**, 62–76 (2017). doi: 10.1016/j.ydbio.2017.02.010; pmid: 28219675
71. J. Holzschuh *et al*., Noradrenergic neurons in the zebrafish hindbrain are induced by retinoic acid and require tfap2a for expression of the neurotransmitter phenotype. *Development* **130**, 5741–5754 (2003). doi: 10.1242/dev.00816; pmid: 14534139
72. R. A. Carter *et al*., A Single-Cell Transcriptional Atlas of the Developing Murine Cerebellum. *Curr. Biol.* **28**, 2910–2920.e2 (2018). doi: 10.1016/j.cub.2018.07.062; pmid: 30220501
73. J. Park *et al*., Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360**, 758–763 (2018). doi: 10.1126/science.aar2131; pmid: 29622724
74. J. Madsen, O. Nielsen, I. Tornøe, L. Thim, U. Holmskov, Tissue localization of human trefoil factors 1, 2, and 3. *J. Histochem. Cytochem.* **55**, 505–513 (2007). doi: 10.1369/jhc.6A7100.2007; pmid: 17242463
75. P. Zhang *et al*., Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. *Cell Rep.* **27**, 1934–1947.e5 (2019). doi: 10.1016/j.celrep.2019.04.052; pmid: 31067475
76. T. F. Griggs *et al*., Rhinovirus C targets ciliated airway epithelial cells. *Respir. Res.* **18**, 84 (2017). doi: 10.1186/s12931-017-0567-0; pmid: 28472984
77. P. A. Reyfman *et al*., Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517–1536 (2019). doi: 10.1164/rccm.201712-24100C; pmid: 30554520
78. S. M. Brierley, T. J. Hibberd, N. J. Spencer, Spinal Afferent Innervation of the Colon and Rectum. *Front. Cell. Neurosci.* **12**, 467 (2018). doi: 10.3389/fncel.2018.00467; pmid: 30564102
79. Y. Liu *et al*., Single-cell RNA-seq reveals the diversity of trophoblast subtypes and patterns of differentiation in the human placenta. *Cell Res.* **28**, 819–832 (2018). doi: 10.1038/s41422-018-0066-y; pmid: 30042384
80. P. S. Zammit *et al*., Pax7 and myogenic progression in skeletal muscle satellite cells. *J. Cell Sci.* **119**, 1824–1832 (2006). doi: 10.1242/jcs.02908; pmid: 16608873
81. T. Schöneberg, J. Meister, A. B. Knierim, A. Schulz, The G protein-coupled receptor GPR34 - The past 20 years of a grownup. *Pharmacol. Ther.* **189**, 71–88 (2018). doi: 10.1016/j.pharmthera.2018.04.008; pmid: 29684466
82. C. Mallika, Q. Guo, J. Y. H. Li, Gbx2 is essential for maintaining thalamic neuron identity and repressing habenular characters in the developing thalamus. *Dev. Biol.* **407**, 26–39 (2015). doi: 10.1016/j.ydbio.2015.08.010; pmid: 26297811
83. J. Cao *et al*., Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018). doi: 10.1126/science.aau0730; pmid: 30166440
84. B. J. Pepe-Mooney *et al*., Single-Cell Analysis of the Liver Epithelium Reveals Dynamic Heterogeneity and an Essential Role for YAP in Homeostasis and Regeneration. *Cell Stem Cell* **25**, 23–38.e8 (2019). doi: 10.1016/j.stem.2019.04.004; pmid: 31080134
85. See supplementary materials.
86. P. A. Reyfman *et al*., Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517–1536 (2019). doi: 10.1164/rccm.201712-24100C; pmid: 30554520
87. A. L. Haber *et al*., A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017). doi: 10.1038/nature24489; pmid: 29144463
88. S. A. MacParland *et al*., Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018). doi: 10.1038/s41467-018-06318-7; pmid: 30348985
89. S. Krasemann *et al*., The TREM2-APOE Pathway Drives the Transcriptional Phenotype of Dysfunctional Microglia in Neurodegenerative Diseases. *Immunity* **47**, 566–581.e9 (2017). doi: 10.1016/j.immuni.2017.08.008; pmid: 28930663
90. T. Kielian, Toll-like receptors in central nervous system glial inflammation and homeostasis. *J. Neurosci. Res.* **83**, 711–730 (2006). doi: 10.1002/jnr.20767; pmid: 16541438
91. H. Kohno, T. Maeda, L. Perusek, E. Pearlman, A. Maeda, CCL3 production by microglial cells modulates disease severity in murine models of retinal degeneration. *J. Immunol.* **192**, 3816–3827 (2014). doi: 10.4049/jimmunol.1301738; pmid: 24639355
92. M. Uittenbogaard, K. K. Baxter, A. Chiaramello, NeuroD6 genomic signature bridging neuronal differentiation to survival via the molecular chaperone network. *J. Neurosci. Res.* **88**, 33–54 (2010). doi: 10.1002/jnr.22182; pmid: 19610105
93. T. Kodama *et al*., Neuronal classification and marker gene identification via single-cell expression profiling of brainstem vestibular neurons subserving cerebellar learning. *J. Neurosci.* **32**, 7819–7831 (2012). doi: 10.1523/JNEUROSCI.0543-12.2012; pmid: 22674258
94. H. Yanai *et al*., Dlk-1, a cell surface antigen on foetal hepatic stem/progenitor cells, is expressed in hepatocellular, colon, pancreas and breast carcinomas at a high frequency. *J. Biochem.* **148**, 85–92 (2010). doi: 10.1093/jb/mvq034; pmid: 20356822
95. D. Zhao *et al*., Derivation and characterization of hepatic progenitor cells from human embryonic stem cells. *PLOS ONE* **4**, e6468 (2009). doi: 10.1371/journal.pone.0006468; pmid: 19649295
96. H. Mutlu-Albayrak *et al*., Identification of SLC22A5 Gene Mutation in a Family with Carnitine Uptake Defect. *Case Rep. Genet.* **2015**, 259627 (2015). doi: 10.1155/2015/259627; pmid: 26075114
97. Z. Huang *et al*., ACSS2 promotes systemic fat storage and utilization through selective regulation of genes involved in lipid metabolism. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9499–E9506 (2018). doi: 10.1073/pnas.1806635115; pmid: 30228117
98. S. C. S. Nagamani, A. Erez, A metabolic link between the urea cycle and cancer cell proliferation. *Mol. Cell. Oncol.* **3**, e1127314 (2016). doi: 10.1080/23723556.2015.1127314; pmid: 27308634
99. H. A. Pliner, J. Shendure, C. Trapnell, Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019). doi: 10.1038/s41592-019-0535-3; pmid: 31501545
100. M. Baron *et al*., A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346–360.e4 (2016). doi: 10.1016/j.cels.2016.08.011; pmid: 27667365
101. T. Abdelaal *et al*., A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019). doi: 10.1186/s13059-019-1795-z; pmid: 31500660
102. I. Korsunsky *et al*., Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019). doi: 10.1038/s41592-019-0619-0; pmid: 31740819

103. L. Castagnaro *et al*., Nkx2-5^{+}islet1^{+} mesenchymal precursors generate distinct spleen stromal cell subsets and participate in restoring stromal network integrity. *Immunity* **38**, 782–791 (2013). doi: 10.1016/j.immuni.2012.12.005; pmid: 23601687
104. A. Oda *et al*., Niche-induced extramedullary hematopoiesis in the spleen is regulated by the transcription factor Tlx1. *Sci. Rep.* **8**, 8308 (2018). doi: 10.1038/s41598-018-26693-x; pmid: 29844356
105. P.-H. Kim, S.-S. Na, B. Lee, J.-H. Kim, J.-Y. Cho, Stanniocalcin 2 enhances mesenchymal stem cell survival by suppressing oxidative stress. *BMB Rep.* **48**, 702–707 (2015). doi: 10.5483/BMBRep.2015.48.12.158; pmid: 26424558
106. A.-C. Villani *et al*., Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356**, eaah4573 (2017). doi: 10.1126/science.aah4573; pmid: 28428369
107. S. Epelman, K. J. Lavine, G. J. Randolph, Origin and functions of tissue macrophages. *Immunity* **41**, 21–35 (2014). doi: 10.1016/j.immuni.2014.06.013; pmid: 25035951
108. D.-M. Popescu *et al*., Decoding human fetal liver haematopoiesis. *Nature* **574**, 365–371 (2019). doi: 10.1038/s41586-019-1652-y; pmid: 31597962
109. Z. Bian *et al*., Deciphering human macrophage development at single-cell resolution. *Nature* **582**, 571–576 (2020). doi: 10.1038/s41586-020-2316-7; pmid: 32499656
110. Immune cell markers (BD Biosciences); www.bdbiosciences.com/us/s/cdmarkers.
111. Immune cell markers (proteintech); www.ptglab.com/products/featured-products/markers-for-immune-cells/.
112. Immune cell marker (Invitrogen eBioscience); www.thermofisher.com/us/en/home/life-science/cell-analysis/cell-analysis-learning-center/cell-analysis-resource-library/ebioscience-resources/human-cd-other-cellular-antigens.html.
113. S. Kitchen *et al*., Ligation of CD4 on human blood monocytes triggers macrophage differentiation (158.8). *J. Immunol.* **186**, 158.8 (2011); www.jimmunol.org/content/186/1_Supplement/158.8.
114. R. G. Domingues *et al*., CD5 expression is regulated during human T-cell activation by alternative polyadenylation, PTBP1, and miR-204. *Eur. J. Immunol.* **46**, 1490–1503 (2016). doi: 10.1002/eji.201545663; pmid: 27005442
115. F. Melo-Gonzalez, M. R. Hepworth, Functional and phenotypic heterogeneity of group 3 innate lymphoid cells. *Immunology* **150**, 265–275 (2017). doi: 10.1111/imm.12697; pmid: 27935637
116. M. Nagasawa *et al*., KLRG1 and NKp46 discriminate subpopulations of human CD117^{+}CRTH2^{-} ILCs biased toward ILC2 or ILC3. *J. Exp. Med.* **216**, 1762–1776 (2019). doi: 10.1084/jem.20190490; pmid: 31201208
117. E. Dzierzak, S. Philipsen, Erythropoiesis: Development and differentiation. *Cold Spring Harb. Perspect. Med.* **3**, a011601 (2013). doi: 10.1101/cshperspect.a011601; pmid: 23545573
118. I. Kobayashi *et al*., Enrichment of hematopoietic stem/progenitor cells in the zebrafish kidney. *Sci. Rep.* **9**, 14205 (2019). doi: 10.1038/s41598-019-50672-5; pmid: 31578390
119. K. J. Knudsen *et al*., ERG promotes the maintenance of hematopoietic stem cells by restricting their differentiation. *Genes Dev.* **29**, 1915–1929 (2015). doi: 10.1101/gad.268409.115; pmid: 26385962
120. B. Psaila *et al*., Single-cell profiling of human megakaryocyte-erythroid progenitors identifies distinct megakaryocyte and erythroid differentiation pathways. *Genome Biol.* **17**, 83 (2016). doi: 10.1186/s13059-016-0939-7; pmid: 27142433
121. K. Kataoka *et al*., Evi1 is essential for hematopoietic stem cell self-renewal, and its expression marks hematopoietic cells with long-term multilineage repopulating activity. *J. Exp. Med.* **208**, 2403–2416 (2011). doi: 10.1084/jem.20110447; pmid: 22084405
122. M. Lapierre *et al*., Expression and role of RIP140/NRIP1 in chronic lymphocytic leukemia. *J. Hematol. Oncol.* **8**, 20 (2015). doi: 10.1186/s13045-015-0116-6; pmid: 25879677
123. B. K. Tusi *et al*., Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018). doi: 10.1038/nature25741; pmid: 29466336
124. D. Pellin *et al*., A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, 2395 (2019). doi: 10.1038/s41467-019-10291-0; pmid: 31160568
125. A. Yumine, S. T. Fraser, D. Sugiyama, Regulation of the embryonic erythropoietic niche: A future perspective. *Blood Res.* **52**, 10–17 (2017). doi: 10.5045/br.2017.52.1.10; pmid: 28401096
126. L. M. Ernst, E. D. Ruchelli, D. S. Huff, *Color Atlas of Fetal and Neonatal Histology* (Springer, 2011).
127. L. I. Gardner, Adrenal function in the premature and newborn. *Pediatrics* **17**, 414–417 (1956). pmid: 13297523
128. T. Kalfa, K. E. McGrath, Analysis of Erythropoiesis Using Imaging Flow Cytometry. *Methods Mol. Biol.* **1698**, 175–192 (2018). doi: 10.1007/978-1-4939-7428-3_10; pmid: 29076090
129. J. E. Merrill, Tumor necrosis factor alpha, interleukin 1 and related cytokines in brain development: Normal and pathological. *Dev. Neurosci.* **14**, 1–10 (1992). doi: 10.1159/000111642; pmid: 1350976
130. S.-Y. Park, M.-J. Kang, J.-S. Han, Interleukin-1 beta promotes neuronal differentiation through the Wnt5a/RhoA/JNK pathway in cortical neural precursor cells. *Mol. Brain* **11**, 39 (2018). doi: 10.1186/s13041-018-0383-6; pmid: 29973222
131. P. R. Ormel *et al*., Microglia innately develop within cerebral organoids. *Nat. Commun.* **9**, 4167 (2018). doi: 10.1038/s41467-018-06684-2; pmid: 30301888
132. T. F. Pais *et al*., The NAD-dependent deacetylase sirtuin 2 is a suppressor of microglial activation and brain inflammation. *EMBO J.* **32**, 2603–2616 (2013). doi: 10.1038/emboj.2013.200; pmid: 24013120
133. J. Zhu *et al*., Aryl Hydrocarbon Receptor Promotes IL-10 Expression in Inflammatory Macrophages Through Src-STAT3 Signaling Pathway. *Front. Immunol.* **9**, 2033 (2018). doi: 10.3389/fimmu.2018.02033; pmid: 30283437
134. A. Lapenna, M. De Palma, C. E. Lewis, Perivascular macrophages in health and disease. *Nat. Rev. Immunol.* **18**, 689–702 (2018). doi: 10.1038/s41577-018-0056-9; pmid: 30127389
135. H. Hu *et al*., Prenatal diagnosis of de novo partial trisomy 18p and partial monosomy 18q recurrent in a family with fatal aortic coarctation. *Gene* **517**, 132–136 (2013). doi: 10.1016/j.gene.2012.12.001; pmid: 23232358
136. L. Sanjurjo *et al*., The human CD5L/AIM-CD36 axis: A novel autophagy inducer in macrophages that modulates inflammatory responses. *Autophagy* **11**, 487–502 (2015). doi: 10.1080/15548627.2015.1017183; pmid: 25713983
137. E. Zigmond *et al*., Infiltrating monocyte-derived macrophages and resident kupffer cells display different ontogeny and functions in acute liver injury. *J. Immunol.* **193**, 344–353 (2014). doi: 10.4049/jimmunol.1400574; pmid: 24890723
138. P. Dutta *et al*., Macrophages retain hematopoietic stem cells in the spleen via VCAM-1. *J. Exp. Med.* **212**, 497–512 (2015). doi: 10.1084/jem.20141642; pmid: 25800955
139. J. Kalucka *et al*., Single-Cell Transcriptome Atlas of Murine Endothelial Cells. *Cell* **180**, 764–779.e20 (2020). doi: 10.1016/j.cell.2020.01.015; pmid: 32059779
140. K. P. Gannon, S. E. McKey, D. E. Stec, H. A. Drummond, Altered myogenic vasoconstriction and regulation of whole kidney blood flow in the ASIC2 knockout mouse. *Am. J. Physiol. Renal Physiol.* **308**, F339–F348 (2015). doi: 10.1152/ajprenal.00572.2014; pmid: 25520010
141. X. Lian *et al*., RXFP1 Receptor Activation by Relaxin-2 Induces Vascular Relaxation in Mice *via* a $G\alpha_{i2}$-Protein/PI3Kß/γ/Nitric Oxide-Coupled Pathway. *Front. Physiol.* **9**, 1234 (2018). doi: 10.3389/fphys.2018.01234; pmid: 30233409
142. B. J. Andreone *et al*., Blood-Brain Barrier Permeability Is Regulated by Lipid Transport-Dependent Suppression of Caveolae-Mediated Transcytosis. *Neuron* **94**, 581–594.e5 (2017). doi: 10.1016/j.neuron.2017.03.043; pmid: 28416077
143. L. Chen *et al*., Tbx1 regulates Vegfr3 and is required for lymphatic vessel development. *J. Cell Biol.* **189**, 417–424 (2010). doi: 10.1083/jcb.200912037; pmid: 20439995
144. W. Tao, R. Moore, Y. Meng, E. R. Smith, X.-X. Xu, Endocytic adaptors Arh and Dab2 control homeostasis of circulatory cholesterol. *J. Lipid Res.* **57**, 809–817 (2016). doi: 10.1194/jlr.M063065; pmid: 27005486
145. A. Furlan, M. Lübke, I. Adameyko, F. Lallemend, P. Ernfors, The transcription factor Hmx1 and growth factor receptor activities control sympathetic neurons diversification. *EMBO J.* **32**, 1613–1625 (2013). doi: 10.1038/emboj.2013.85; pmid: 23591430
146. Y.-C. Wang *et al*., Homeodomain transcription factor NKX2.2 functions in immature cells to control enteroendocrine differentiation and is expressed in gastrointestinal neuroendocrine tumors. *Endocr. Relat. Cancer* **16**, 267–279 (2009). doi: 10.1677/ERC-08-0127; pmid: 18987169
147. Y. Kobayashi, P. R. Tata, Pulmonary Neuroendocrine Cells: Sensors and Sentinels of the Lung. *Dev. Cell* **45**, 425–426 (2018). doi: 10.1016/j.devcel.2018.05.009; pmid: 29787707
148. C. Li *et al*., Apc deficiency alters pulmonary epithelial cell fate and inhibits Nkx2.1 via triggering TGF-beta signaling. *Dev. Biol.* **378**, 13–24 (2013). doi: 10.1016/j.ydbio.2013.03.018; pmid: 23562608
149. A. J. Churchill *et al*., Genetic evidence that Nkx2.2 acts primarily downstream of Neurog3 in pancreatic endocrine lineage development. *eLife* **6**, e20010 (2017). doi: 10.7554/eLife.20010; pmid: 28071588
150. L. Arnes, J. T. Hill, S. Gross, M. A. Magnuson, L. Sussel, Ghrelin expression in the mouse pancreas defines a unique multipotent progenitor population. *PLOS ONE* **7**, e52026 (2012). doi: 10.1371/journal.pone.0052026; pmid: 23251675
151. F. M. Gribble, F. Reimann, Enteroendocrine Cells: Chemosensors in the Intestinal Epithelium. *Annu. Rev. Physiol.* **78**, 277–299 (2016). doi: 10.1146/annurev-physiol-021115-105439; pmid: 26442437
152. C. Ramond *et al*., Understanding human fetal pancreas development using subpopulation sorting, RNA sequencing and single-cell profiling. *Development* **145**, dev165480 (2018). doi: 10.1242/dev.165480; pmid: 30042179
153. M. Volante *et al*., Ghrelin expression in fetal, infant, and adult human lung. *J. Histochem. Cytochem.* **50**, 1013–1021 (2002). doi: 10.1177/002215540205000803; pmid: 12133904
154. P. LeSimple *et al*., Trefoil factor family 3 peptide promotes human airway epithelial ciliated cell differentiation. *Am. J. Respir. Cell Mol. Biol.* **36**, 296–303 (2007). doi: 10.1165/rcmb.2006-0270OC; pmid: 17008636
155. Y. Taché, W. Marki, J. Rivier, W. Vale, M. Brown, Central nervous system inhibition of gastric secretion in the rat by gastrin-releasing peptide, a mammalian bombesin. *Gastroenterology* **81**, 298–302 (1981). doi: 10.1016/S0016-5085(81)80061-3; pmid: 7239137
156. R. Kurotani *et al*., Role of secretoglobin 3A2 in lung development. *Am. J. Respir. Crit. Care Med.* **178**, 389–398 (2008). doi: 10.1164/rccm.200707-1104OC; pmid: 18535256
157. M. Hochane *et al*., Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLOS Biol.* **17**, e3000152 (2019). doi: 10.1371/journal.pbio.3000152; pmid: 30789893
158. B. Dekel *et al*., Multiple imprinted and stemness genes provide a link between normal and tumor progenitor cells of the developing human kidney. *Cancer Res.* **66**, 6040–6049 (2006). doi: 10.1158/0008-5472.CAN-05-4528; pmid: 16778176
159. T. Moriguchi *et al*., MafB is essential for renal development and F4/80 expression in macrophages. *Mol. Cell. Biol.* **26**, 5715–5727 (2006). doi: 10.1128/MCB.00001-06; pmid: 16847325
160. Y. Maezawa *et al*., Loss of the podocyte-expressed transcription factor Tcf21/Pod1 results in podocyte differentiation defects and FSGS. *J. Am. Soc. Nephrol.* **25**, 2459–2470 (2014). doi: 10.1681/ASN.2013121307; pmid: 24904088
161. S. S. Marable, E. Chung, M. Adam, S. S. Potter, J.-S. Park, Hnf4a deletion in the mouse kidney phenocopies Fanconi renotubular syndrome. *JCI Insight* **3**, e97497 (2018). doi: 10.1172/jci.insight.97497; pmid: 30046000
162. R. O'Rahilly, Early human development and the chief sources of information on staged human embryos. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **9**, 273–280 (1979). doi: 10.1016/0028-2243(79)90068-6; pmid: 400868
163. K. Theiler, *The House Mouse: Atlas of Embryonic Development* (Springer, 2013).
164. K. Reemst, S. C. Noctor, P. J. Lucassen, E. M. Hol, The Indispensable Roles of Microglia and Astrocytes during Brain Development. *Front. Hum. Neurosci.* **10**, 566 (2016). doi: 10.3389/fnhum.2016.00566; pmid: 27877121
165. J. Dai, K. K. Bercury, J. T. Ahrendsen, W. B. Macklin, Olig1 function is required for oligodendrocyte differentiation in the mouse brain. *J. Neurosci.* **35**, 4386–4402 (2015). doi: 10.1523/JNEUROSCI.4962-14.2015; pmid: 25762682
166. J. See *et al*., Oligodendrocyte maturation is inhibited by bone morphogenetic protein. *Mol. Cell. Neurosci.* **26**, 481–492 (2004). doi: 10.1016/j.mcn.2004.04.004; pmid: 15276151
167. K. Komorowska *et al*., Hepatic Leukemia Factor Maintains Quiescence of Hematopoietic Stem Cells and Protects the Stem Cell Pool during Regeneration. *Cell Rep.* **21**, 3514–3523 (2017). doi: 10.1016/j.celrep.2017.11.084; pmid: 29262330
168. H. Huang, Y. Li, B. Liu, Transcriptional regulation of mast cell and basophil lineage commitment. *Semin. Immunopathol.* **38**, 539–548 (2016). doi: 10.1007/s00281-016-0562-4; pmid: 27126100
169. J. Medvedovic, A. Ebert, H. Tagoh, M. Busslinger, Pax5: A master regulator of B cell development and leukemogenesis. *Adv. Immunol.* **111**, 179–206 (2011). doi: 10.1016/B978-0-12-385991-4.00005-2; pmid: 21970955

170. C. Cantù *et al*., Sox6 enhances erythroid differentiation in human erythroid progenitors. *Blood* **117**, 3669–3679 (2011). doi: 10.1182/blood-2010-04-282350; pmid: 21263153
171. Y. Lavin, A. Mortha, A. Rahman, M. Merad, Regulation of macrophage development and function in peripheral tissues. *Nat. Rev. Immunol.* **15**, 731–744 (2015). doi: 10.1038/nri3920; pmid: 26603899
172. X. Qiu *et al*., Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017). doi: 10.1038/nmeth.4402; pmid: 28825705
173. A. Buttgereit *et al*., Sall1 is a transcriptional regulator defining microglia identity and function. *Nat. Immunol.* **17**, 1397–1406 (2016). doi: 10.1038/ni.3585; pmid: 27776109
174. I. R. Holtman, D. Skola, C. K. Glass, Transcriptional control of microglia phenotypes in health and disease. *J. Clin. Invest.* **127**, 3220–3229 (2017). doi: 10.1172/JCI90604; pmid: 28758903
175. P. Mazzarello, A unifying concept: The history of cell theory. *Nat. Cell Biol.* **1**, E13–E15 (1999). doi: 10.1038/8964; pmid: 10559875
176. M. Asp *et al*., A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell* **179**, 1647–1660.e19 (2019). doi: 10.1016/j.cell.2019.11.025; pmid: 31835037
177. D. E. Wagner *et al*., Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018). doi: 10.1126/science.aar4362; pmid: 29700229
178. J. A. Briggs *et al*., The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018). doi: 10.1126/science.aar5780; pmid: 29700227
179. N. Karaiskos *et al*., The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017). doi: 10.1126/science.aan3235; pmid: 28860209
180. P. Blakeley *et al*., Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3613 (2015). doi: 10.1242/dev.131235; pmid: 26487783
181. B. Pijuan-Sala *et al*., A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019). doi: 10.1038/s41586-019-0933-9; pmid: 30787436
182. P. Wang *et al*., Dissecting the Global Dynamic Molecular Profiles of Human Fetal Kidney Development by Single-Cell RNA Sequencing. *Cell Rep.* **24**, 3554–3567.e3 (2018). doi: 10.1016/j.celrep.2018.08.056; pmid: 30257215
183. J. M. Segal *et al*., Single cell analysis of human foetal liver captures the transcriptional profile of hepatobiliary hybrid progenitors. *Nat. Commun.* **10**, 3350 (2019). doi: 10.1038/s41467-019-11266-x; pmid: 31350390
184. S. Gao *et al*., Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* **20**, 721–734 (2018). doi: 10.1038/s41556-018-0105-4; pmid: 29802404
185. J. S. Banerji, K. M. Ramani, A. Devasia, Extramedullary hematopoiesis in the adrenal: Case report and review of literature. *Can. Urol. Assoc. J.* **7**, E436–E438 (2013). doi: 10.5489/cuaj.1389; pmid: 23826059
186. H. Al-Thani, M. Al-Sulaiti, G. El-Mabrok, A. Tabeb, A. El-Menyar, Adrenal extramedullary hematopoiesis associated with beta-thalassemia trait in an adult woman: A case report and review of literature. *Int. J. Surg. Case Rep.* **24**, 83–87 (2016). doi: 10.1016/j.ijscr.2016.05.024; pmid: 27232291
187. J. Cao, J. Shendure, sci-RNA-seq3, protocols.io (2019); https://dx.doi.org/10.17504/protocols.io.9yih7ue.
188. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019). doi: 10.1016/J.CELS.2018.11.005; pmid: 30954476
189. P. J. Thul, C. Lindskog, The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018). doi: 10.1002/pro.3307; pmid: 28940711
190. S. Aibar *et al*., SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017). doi: 10.1038/nmeth.4463; pmid: 28991892
191. S. Amini *et al*., Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014). doi: 10.1038/ng.3119; pmid: 25326703
192. J. Cao, sci-RNA-seq3 processing pipeline, Zenodo (2020); http://doi.org/10.5281/zenodo.4013713.

ACKNOWLEDGMENTS

We thank past and present members of the Shendure and Trapnell laboratories, A. Adey, the Brotman Baty Institute Advanced Technology Lab, and the University of Washington Birth Defects Research Laboratory. We thank C. Spurrell and D. Ahrendsen for reviewing the sci-RNA-seq3 protocol. **Funding:** M.S. was supported by grants from the Deutsche Forschungsgemeinschaft (SP1532/3-1, 4-1, and 5-1). Aspects of this work were supported by funding from the Brotman Baty Institute for Precision Medicine, the Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to J.S. and C.T.), the Chan Zuckerberg Initiative (to C.T.), and the NIH (DK119285 to J.P.; HD000836 to I.A.G.). J.S. is an investigator of the Howard Hughes Medical Institute. **Author contributions:** J.C. developed techniques and performed sci-RNA-seq3 experiments with assistance from R.M.D., R.B., F.Z., and F.J.S.; D.R.O. and M.D. performed tissue collection, nuclei extraction, and immunohistochemistry staining under the supervision of D.D. and I.A.G.; K.A.A. performed genotyping; P.D.K. and J.P. performed erythropoiesis validation in mice; J.C. performed computational analyses with assistance from H.A.P. and M.S.; M.A.Z. developed the website with assistance from J.C.; J.C. and J.S. wrote the manuscript with input from all coauthors; and J.S. and C.T. supervised the project. **Competing interests:** F.Z. and F.J.S. declare competing financial interests in the form of stock ownership and paid employment by Illumina, Inc. J.S. has competing financial interests (paid consulting and/or equity) with Guardant Health, Maze Therapeutics, Camp4 Therapeutics, Nanostring, Phase Genomics, Adaptive Biotechnologies, and Stratos Genomics. One or more embodiments of one or more patents and patent applications filed by Illumina and the University of Washington may encompass the methods, reagents, and data disclosed in this manuscript. The authors declare no other competing interests. **Data and materials availability**: A detailed version of the sci-RNA-seq3 protocol is available on protocols.io (*187*). All methods for making the transposase complexes are described in (*191*), and Tn5 transposase is available from Lucigen; however, Illumina will provide transposase complexes in response to reasonable requests from the scientific community, subject to a material transfer agreement. A demultiplexing script and tutorial are provided on Zenodo (*192*). Raw data are provided at dbGaP (accession no. phs002003.v1.p1). Supplementary files S1 to S7 and table S5 (which is also provided in the supplementary materials) are provided at GEO (accession no. GSE156793). Garnett models generated from these data are also posted to our website, where they can broadly be used for the automated cell type classification of single-cell data from diverse organs. An interactive version of the full dataset facilitates the exploration of these data by tissue, cell type, or gene (descartes.brotmanbaty.org).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/eaba7721/suppl/DC1
Supplementary Materials and Methods
Figs. S1 to S27
Tables S1 to S16
References (*193*–*207*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

3 January 2020; accepted 10 September 2020
10.1126/science.aba7721

RESEARCH ARTICLE SUMMARY

HUMAN GENOMICS

A human cell atlas of fetal chromatin accessibility

Silvia Domcke*, Andrew J. Hill*, Riza M. Daza*, Junyue Cao, Diana R. O'Day, Hannah A. Pliner, Kimberly A. Aldinger, Dmitry Pokholok, Fan Zhang, Jennifer H. Milbank, Michael A. Zager, Ian A. Glass, Frank J. Steemers, Dan Doherty, Cole Trapnell†, Darren A. Cusanovich†, Jay Shendure†

INTRODUCTION: In recent years, the single-cell genomics field has made incredible progress toward disentangling the cellular heterogeneity of human tissues. However, the overwhelming majority of effort has been focused on single-cell gene expression rather than the chromatin landscape that shapes and is shaped by gene expression. Toward advancing our understanding of the regulatory programs that underlie human cell types, we set out to generate single-cell atlases of both chromatin accessibility (this study) and gene expression (Cao *et al.*, this issue) from a broad range of human fetal tissues.

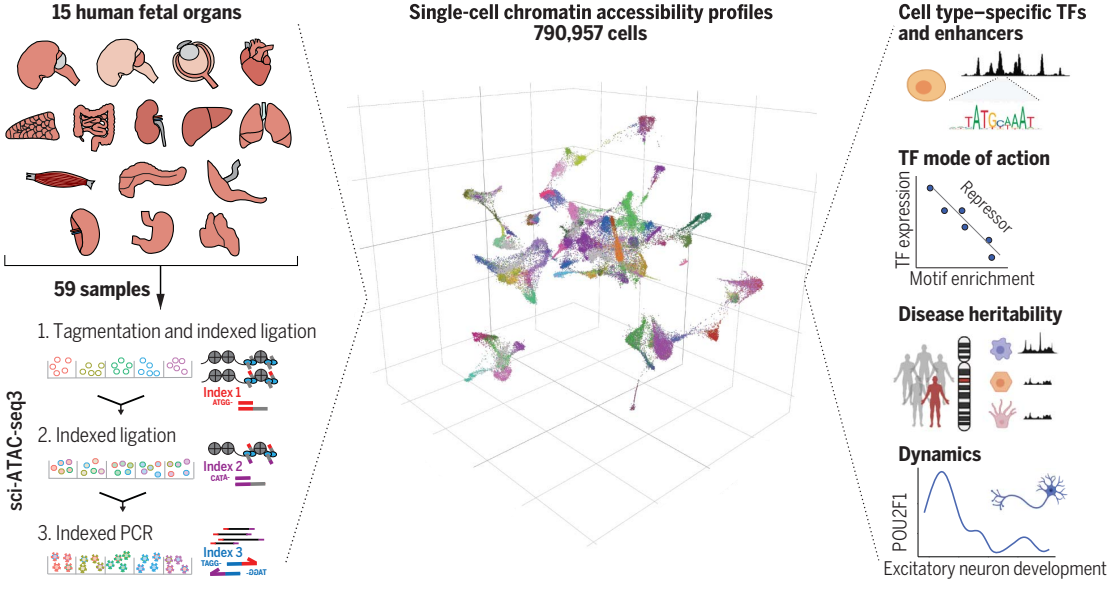
RATIONALE: Regions of accessible chromatin in our genome, such as enhancers, play key roles in the determination and maintenance of cell fates. Accessible regions are also markedly enriched for genetic variation that contributes to common disease heritability. The vast majority of chromatin accessibility data collected to date lacks single-cell resolution, limiting our ability to infer patterns such as which cell types are most relevant to each common disease. We previously demonstrated single-cell profiling of chromatin accessibility using combinatorial indexing, based on two rounds of in situ molecular barcoding. Here, we describe an improved assay that uses three levels of combinatorial indexing and does not rely on custom reagents. The method, sci-ATAC-seq3, reduces costs and opens the door to the scales necessary for generating a human cell atlas of chromatin accessibility.

RESULTS: We applied sci-ATAC-seq3 to 59 human fetal samples ranging from 89 to 125 days in estimated postconceptual age and representing 15 organs, altogether obtaining high-quality chromatin accessibility profiles from ~800,000 single cells. Gene expression data collected on an overlapping set of tissues were leveraged to annotate cell types. We asked which transcription factor (TF) motifs found in the accessible sites of each cell best explain its cell type affiliation, revealing both known and potentially previously unknown regulators of cell fate specification and/or maintenance. Many TFs could be putatively assigned as activators or repressors depending on whether their expression and the accessibility of their cognate motif were positively or negatively correlated across cell types. Comparing chromatin accessibility from cell types that appear in multiple tissues revealed that whereas blood cell types are highly similar across organs, endothelial cells exhibit organ-specific chromatin accessibility, which appears to be controlled combinatorially by several TFs with overlapping expression patterns. We leveraged our master set of 1.05 million accessible sites, spanning 532 Mb or 17% of the reference human genome, to score cell type–specific links between candidate enhancers and genes based on coaccessibility, to detect cell type–specific enrichment of heritability for specific common human diseases, and to identify genetic variants affecting chromatin accessibility in cis. Comparisons with chromatin accessibility in corresponding adult tissues allowed us to identify fetal-specific cell subtypes and nominate POU2F1 as a potential regulator of excitatory neuron development.

CONCLUSION: Sci-ATAC-seq3 adds to a growing repertoire of single-cell methods that use combinatorial indexing, a technical paradigm whose advantages include exponential scaling and greater range to profile diverse aspects of single-cell biology. We anticipate that the intersection of single-cell chromatin accessibility and gene expression will critically accelerate the field's long-term goal of establishing a deep, predictive understanding of gene regulation. An interactive website facilitates the exploration of these freely available data by tissue, cell type, locus, or motif (descartes.brotmanbaty.org). ■

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: darrenc@email.arizona.edu (D.A.C.); coletrap@uw.edu (C.T.); shendure@uw.edu (J.S.)
Cite this article as S. Domcke *et al.*, *Science* **370**, eaba7612 (2020). DOI: 10.1126/science.aba7612

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.aba7612



A human cell atlas of fetal chromatin accessibility enables the exploration of in vivo gene regulation across diverse cell types. We devised a three-level combinatorial indexing assay (sci-ATAC-seq3) and profiled chromatin accessibility in ~800,000 single cells from 15 fetal organs. This rich resource enables, for example, identification of cell type–specific regulatory elements and TFs, classification of TFs into activators and repressors, and quantification of cell type–specific enrichments of complex trait heritability, as well as chromatin accessibility dynamics.

RESEARCH ARTICLE

HUMAN GENOMICS

A human cell atlas of fetal chromatin accessibility

Silvia Domcke[1]*, Andrew J. Hill[1]*, Riza M. Daza[1]*, Junyue Cao[1], Diana R. O'Day[2], Hannah A. Pliner[3], Kimberly A. Aldinger[2,4], Dmitry Pokholok[5], Fan Zhang[5], Jennifer H. Milbank[1], Michael A. Zager[3,6], Ian A. Glass[2,3,4], Frank J. Steemers[5], Dan Doherty[2,3,4], Cole Trapnell[1,3,7]†, Darren A. Cusanovich[1,8,9]†, Jay Shendure[1,3,7,10]†

The chromatin landscape underlying the specification of human cell types is of fundamental interest. We generated human cell atlases of chromatin accessibility and gene expression in fetal tissues. For chromatin accessibility, we devised a three-level combinatorial indexing assay and applied it to 53 samples representing 15 organs, profiling ~800,000 single cells. We leveraged cell types defined by gene expression to annotate these data and cataloged hundreds of thousands of candidate regulatory elements that exhibit cell type–specific chromatin accessibility. We investigated the properties of lineage-specific transcription factors (such as POU2F1 in neurons), organ-specific specializations of broadly distributed cell types (such as blood and endothelial), and cell type–specific enrichments of complex trait heritability. These data represent a rich resource for the exploration of in vivo human gene regulation in diverse tissues and cell types.

In recent years, the single-cell genomics field has made incredible progress toward disentangling the cellular heterogeneity of human tissues. However, the overwhelming majority of effort has been focused on single-cell gene expression, with far fewer investigations of the chromatin landscape that shapes and is shaped by gene expression. This is in part because of a relative paucity of scalable methods for profiling chromatin accessibility, transcription factor (TF) binding, and/or histones at single-cell resolution.

The single-cell combinatorial indexing ("sci-") (*1*) framework involves the splitting and pooling of cells or nuclei to wells in which molecular barcodes are introduced in situ to a species of interest at each round. Through successive rounds of in situ molecular barcoding, species within the same cell are concordantly labeled with a distinct combination of barcodes. Sci- assays have been developed for profiling chromatin accessibility [sci-ATAC-seq (ATAC-seq, assay for transposase-accessible chromatin with high-throughput sequencing)], gene expression [sci-RNA-seq (RNA-seq, RNA-sequencing)], nuclear architecture, genome sequence, methylation, histone marks and other phenomena, as well as sci- co-assays—for example, for profiling chromatin accessibility and gene expression jointly (*1*–*12*) ["CoBatch," "Split-seq," "Paired-seq," and "dscATAC-seq" also effectively rely on single-cell combinatorial indexing (*8*–*10*, *12*)]. Although we and others have profiled chromatin accessibility in >100,000 mammalian cells (*9*, *12*, *13*), the methods used require custom-loading of the Tn5 enzyme with barcoded adapters and/or are limited to 10^4 to 10^5 cells per experiment by collisions—cells receiving the same combination of barcodes.

We developed an improved assay for single-cell profiling of chromatin accessibility that both uses three levels of combinatorial indexing and, in contrast with previous iterations of sci-ATAC-seq and related methods (*1*, *6*, *9*, *12*), does not rely on molecularly barcoded Tn5 complexes (sci-ATAC-seq3) (Fig. 1A and fig. S1A). Rather, the first two rounds of indexing are achieved through ligation to either end of the conventional, uniformly loaded Tn5 transposase complex (standard Nextera), whereas the final round of indexing remains through polymerase chain reaction (PCR). Relative to two-level sci-ATAC-seq but similar to sci-RNA-seq3, sci-ATAC-seq3 reduces the per-cell cost of library preparation (fig. S1B) as well as the rate of collisions (fig. S1, C and D), opening the door to experiments on the scale of 10^6 cells. This protocol no longer requires cell sorting, and we also optimized ligase and polymerase choice, kinase concentration, and oligo designs and concentrations to maximize the number of fragments recovered from each cell. While maintaining an enrichment in accessible regions, we made the explicit choice to maximize complexity at the expense of specificity for accessible sites (Fig. 1B and fig. S1, E to G). In particular, we found that the fixation conditions could be tuned to adjust the sensitivity (complexity) versus specificity (enrichment in accessible sites) of the assay (fig. S1H).

As one step toward a comprehensive cell atlas of human development (*14*), we set out to generate single-cell atlases of both gene expression and chromatin accessibility using diverse human tissues obtained during mid-gestation [DESCARTES, Developmental Single Cell Atlas of gene Regulation and Expression; descartes.brotmanbaty.org (*15*)]. For chromatin accessibility, we applied sci-ATAC-seq3 to 59 fetal samples representing 15 organs, altogether profiling 1.6 million cells (Fig. 1C). We also describe profiling of gene expression in 5 million cells from the same organs, using an overlapping set of samples (*16*). The profiled organs span diverse systems. However, some systems were not accessible; bone marrow, bone, gonads, and skin are notably absent.

The rapid and uniform processing of heterogeneous fetal tissues presents a challenge. We developed a method for extracting nuclei directly from cryopreserved tissues that works across a variety of tissue types and produces homogenates suitable for both sci-ATAC-seq3 and sci-RNA-seq3. For sci-ATAC-seq3, we used tissue samples obtained from 23 fetuses ranging from 89 to 125 days in estimated post-conceptual age (Fig. 1, D and E, and table S1). All samples were karyotypically normal. Samples were processed in three batches; a mix of the same sentinel human fetal brain tissue and a mouse suspension cell line was included in each experiment to control for batch effects and estimate collision rates.

We sequenced sci-ATAC-seq3 libraries from the three experimental batches across five Illumina NovaSeq 6000 sequencing runs, generating just over 110 billion reads (55 billion read pairs). We compared these data at the tissue level, before splitting into single cells, against single-ended ENCODE deoxyribonuclease-sequencing (DNase-seq) data (fig. S2A) (*17*). Although sci-ATAC-seq3 data were somewhat less enriched in peaks (median reads in peaks: 29% for sci-ATAC-seq3; 35% for ENCODE DNase-seq) (fig. S2B), samples from the same tissue were comparably correlated for the two assays (median Spearman correlation: 0.93 for two samples from the same tissue for sci-ATAC-seq3; 0.91 for DNase-seq), with greater technical reproducibility for sci-ATAC-seq3 (median Spearman correlation: 0.95) (fig. S2C). Furthermore, samples clustered into their respective tissues from these aggregate profiles, whether analyzing the sci-ATAC-seq3 samples alone (Fig. 1F) or the sci-ATAC-seq3 and DNase-seq samples together (fig. S2D).

Splitting reads by sci- barcodes, we identified 1,568,018 cells (table S1), and from the barnyard control, we estimated collision rates of 1 to 4% for the three experiments (fig. S2E) (*18*). We observed no obvious batch effects (fig. S2F)

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [2]Department of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA. [3]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA. [4]Center for Integrative Brain Research, Seattle Children's Research Institute, Seattle, WA, USA. [5]Illumina, San Diego, CA, USA. [6]Center for Data Visualization, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [7]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA. [8]Department of Cellular and Molecular Medicine, University of Arizona, Tucson, AZ, USA. [9]Asthma and Airway Disease Research Center, University of Arizona, Tucson, AZ, USA. [10]Howard Hughes Medical Institute, Seattle, WA, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: darrenc@email.arizona.edu (D.A.C.); coletrap@uw.edu (C.T.); shendure@uw.edu (J.S.)

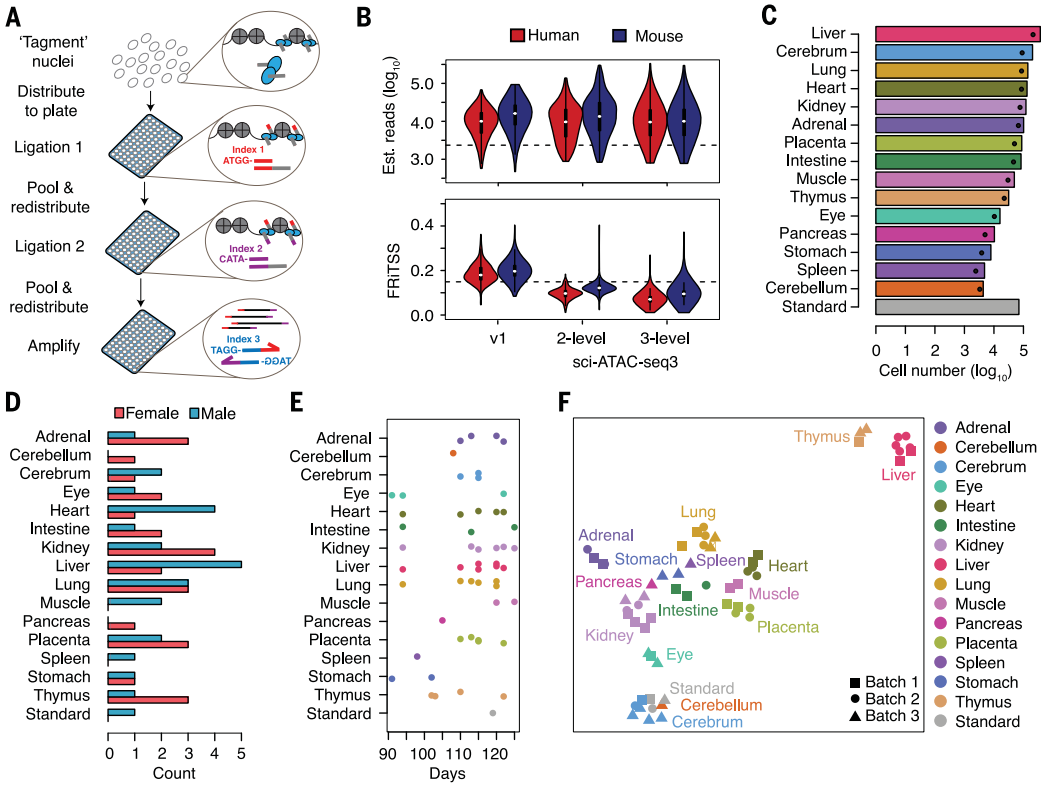


Fig. 1. Design of three-level sci-ATAC-seq and application to chromatin accessibility profiling of 1.6 million cells from 59 fetal samples. (**A**) Schematic of sci-ATAC-seq3. Nuclei are tagmented with Tn5 transposase in bulk. The first two rounds of indexing are achieved with successive ligations to each end of the Tn5 transposase complex, and the third round is achieved with PCR. (**B**) Comparison of complexity and specificity achieved with different versions of the sci-ATAC-seq protocol in mixing experiments of mouse and human suspension cell lines. The estimated total nonduplicate reads ("complexity") for each cell were calculated with Picard and are displayed as violin plots on a log10 scale (*115*). The fraction of reads in TSSs (FRiTSS) was calculated for each cell in the same experiments (bottom). Reads within 500 bp of a Gencode TSS were considered within the TSS. v1: species mixing experiment by using our previously published two-level sci-ATAC-seq protocol (*13*); 2-level: two-level version of the new protocol with simultaneous ligations; and 3-level: three-level version of the new protocol. (**C**) Barplot showing number of cells profiled per organ (log10 scale). Dots indicate the number of cells remaining after QC filtering procedures. Standard: sentinel tissue (trisomy 18 cerebrum) was included in all three experiments. (**D**) Barplot showing the distribution of sexes for samples corresponding to each organ. (**E**) Stripchart showing the estimated post-conceptual age of each sample. Samples are arranged by organ and slightly jittered to avoid overplotting. (**F**) UMAP visualization of aggregated chromatin accessibility profiles of single cells from each of the samples, colored by organ. Normalized accessibility at a master set of peaks was quantified for each "pseudobulk" sample and used as an input to UMAP. Shapes indicate the processing batch of each sample.

and dropped three samples on account of poor nucleosomal banding of their fragment size distribution (fig. S2G) and a further two samples that captured few cells. For the remaining samples, we observed a median of 5742 nonduplicate reads per cell (fig. S2H) and estimate that we sequenced a median of 88% of all nonduplicate reads per cell in these sci-ATAC-seq3 libraries (fig. S2I).

We identified peaks of accessibility on a tissue-by-tissue basis and then merged these to generate a master set of 1.05 million sites (data file S1). We filtered out lower-quality cells, which left 790,957 single-cell chromatin accessibility profiles from 53 fetal samples (data file S2). The total number of high-quality cells per tissue ranged from 2421 for spleen to 211,450 for liver (Fig. 1C). The median number of nonduplicate fragments per cell for this set is 6042, with a median of 49% overlapping the master set of accessible sites and 19% falling near a transcription start site (TSS) (±1 kb). We subjected high-quality cells to latent semantic indexing (*19*, *20*), linear correction (*21*), and Louvain clustering, initially obtaining 172 clusters across all tissues. We further reduced the dimensionality of each tissue dataset using Uniform Manifold Approximation and Projection (UMAP) (*22*).

Annotating cell types

The annotation of cell types in scATAC-seq (sc, single cell) datasets can be simplified by leveraging scRNA-seq datasets (*13*, *23*–*25*). In order to partially automate cell type annotations for our sci-ATAC-seq data, we first annotated cell types within our sci-RNA-seq data for the same tissues (*16*). Second, we computed gene-level accessibility scores for our sci-ATAC-seq data, aggregating the number of transposition events falling within gene bodies extended by 2 kb upstream of their TSS. Third, we used the gene-by-cell matrices for each data type as input to an approach for finding likely correspondences between clusters on the basis of non-negative least squares (NNLS) regression (*26*), effectively resulting in a "lift-over" set of automated annotations for our sci-ATAC-seq clusters. Last, we manually reviewed these automated annotations by examining pileups around marker genes for each cell type within each tissue, making modifications to assigned labels as deemed necessary (Fig. 2A and fig. S3A). Although other approaches have shown considerable promise for multimodal integration of single-cell data (*23*), we found this cluster-to-cluster NNLS method (*26*) sufficient for our purposes here and much less computationally intensive.

Altogether, we were able to annotate 150 of the 172 clusters (87%), or 163 of 172 (95%) if we include lower-confidence labels. Some clusters received the same annotation within the same tissue and were merged, resulting in 124 annotations across all tissues. Of these, some annotations were present across multiple tissues (Fig. 2B). Collapsing across tissues resulted in 54 distinct cell type annotations that map 1:1 to "main cell type" annotations made in our sci-RNA-seq dataset (or 59 if we include lower-confidence labels and 1:2 mappings) (Fig. 2B). Many of the sci-RNA-seq cell types that were not found in the sci-ATAC-seq data at this level of resolution are small clusters that may not have been sufficiently sampled to be detectable, owing to the lower number of cells profiled here [~4 million RNA (*16*) versus ~800,000 ATAC high-quality cells] (fig. S3B). However, most of the nine sci-ATAC-seq clusters that remained fully unannotated appear to be due to unfiltered doublets because they are characterized by accessibility in marker genes for multiple

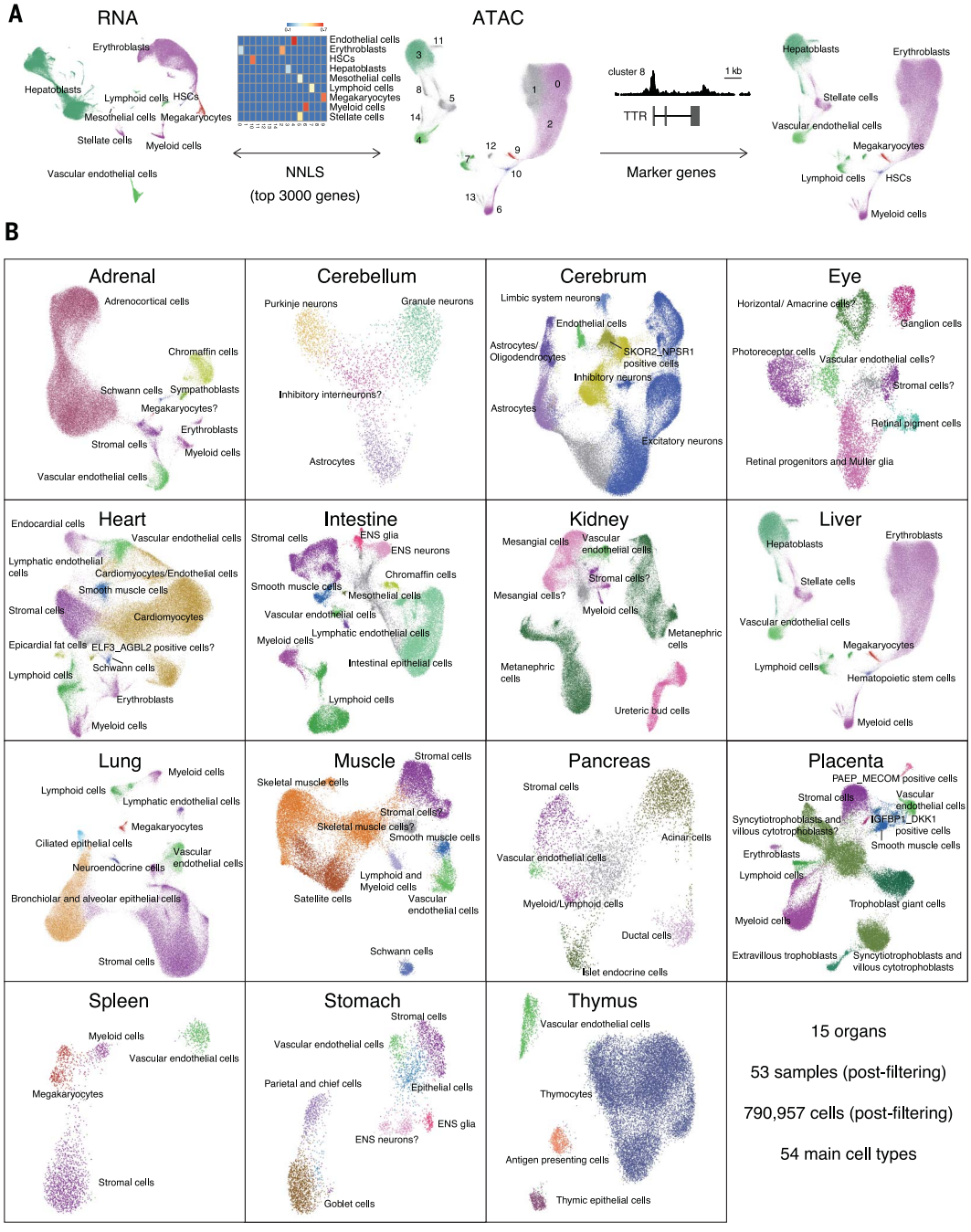


Fig. 2. Identifying cell types across 15 human organs. (**A**) Summary of annotation strategy. (Left) Cell types were first annotated in sci-RNA-seq data (*16*) gathered from matching tissues according to marker gene expression. (Middle) Louvain clusters were identified in sci-ATAC-seq data for each tissue. Next, gene-level accessibility scores were calculated for each of these clusters and matched to RNA clusters on the basis of NNLS regression, in some cases leading to merging of Louvain clusters. (Right) These first-pass automated annotations were refined by manually reviewing the cluster-specific accessibility landscape around marker genes—for example, initially unannotated cluster 8 exhibited specific accessibility at the *TTR* locus—and was therefore merged with cluster 3 (hepatoblasts). (**B**) UMAP visualization and annotation of 790,957 cells profiled across 15 organs. The colors correspond to the 54 main cell types that were identified across the different organs.

adjacent cell types in the UMAP representation (fig. S3A).

The nature of ATAC-seq data allows sexing of cells on the basis of Y chromosome reads. In the placenta in particular, we found that three cell types—*PAEP*$^+$, *MECOM*$^+$ and *IGFBP*$^+$, *DKK*$^+$ cells (both initially unannotated in the RNA data, although the labels readily lifted over to clusters in the ATAC data), as well as placental lymphoid cells—exhibited a significantly lower ratio of Y chromosome–derived reads in tissues derived from male fetuses (fig. S3C). Consistent with what is known about *PAEP* (glycodelin) and *IGFBP1*, these cell types likely correspond to maternally derived endometrial epithelial and decidualized stromal cells, respectively (*27*). This was confirmed with genotype inference with souporcell (*28*), which additionally identifies a subgroup of placental myeloid cells as likely to be of maternal origin (fig. S3D).

Identifying cell type–specific TFs

We next sought to integrate and compare chromatin accessibility in cell types across all 15 organs. To mitigate the effects of gross differences in the numbers of cells per organ and/or cell type, we randomly sampled 800 cells per cell type per organ (including unannotated clusters; in cases in which fewer than 800 cells of a given cell type were represented in a given organ, all cells were taken), and we performed UMAP visualization (Fig. 3A). Reassuringly, cell types represented in multiple organs clustered together—for example, stromal cells (nine organs), endothelial cells (13 organs), lymphoid cells (seven organs), and myeloid cells (10 organs)—rather than by batch or individual (fig. S4). Developmentally and functionally related cell types also colocalized, such as diverse blood cells, secretory cells, peripheral nervous system neurons, and central nervous system neurons.

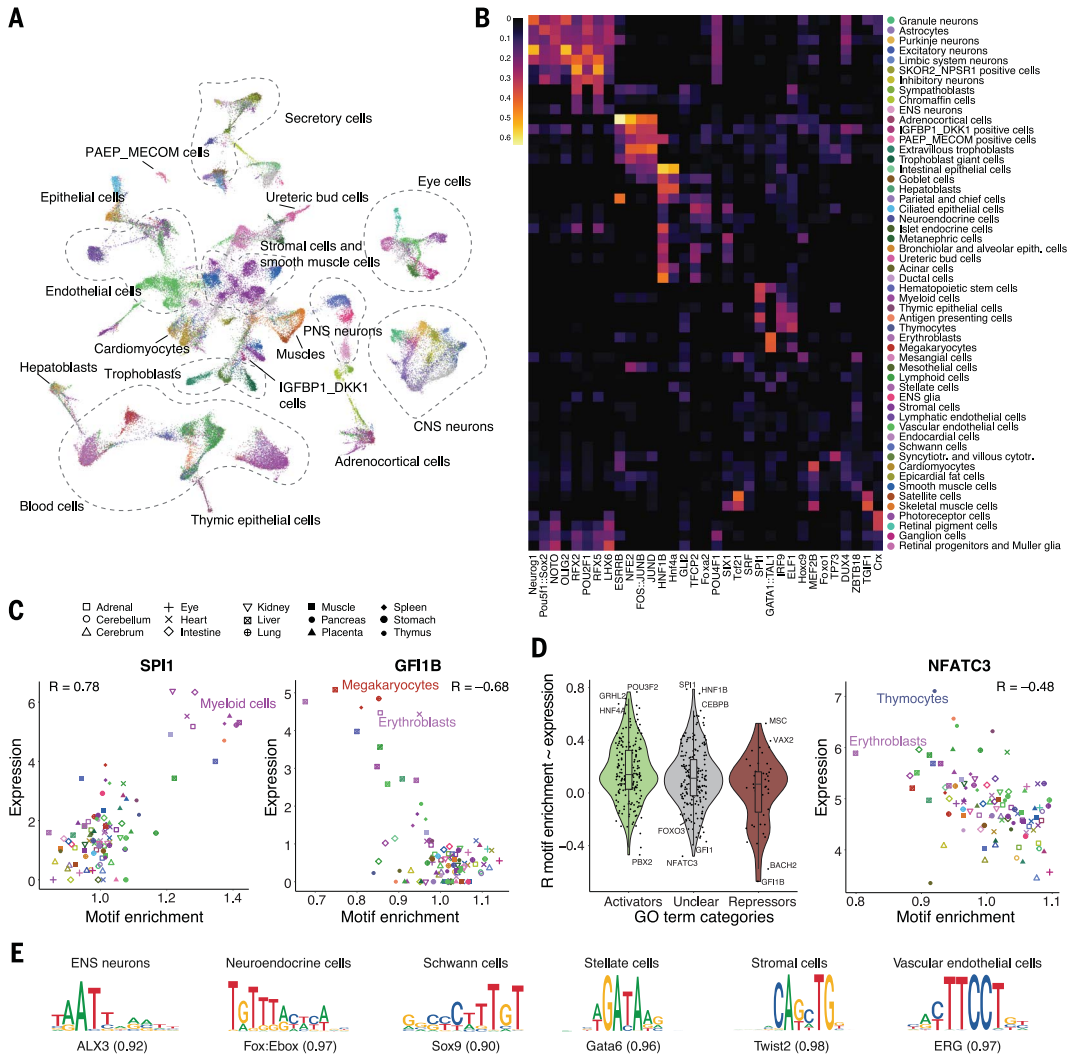


Fig. 3. Identifying key TF regulators of cell type–specific chromatin accessibility and their modes of action. (**A**) Combined UMAP of the entire dataset subsampled to a maximum of 800 cells per initial cluster ID. Cells are colored by 54 main cell types as in (B). Groups of related cell types are circled. (**B**) Fold-change of the top enriched TF motif in cell type–specific peaks for all 54 main cell types. Cell types (rows) are ordered by hierarchical clustering of the motif enrichment matrix (log10-scaled fold-change of the mean motif occurrence in peaks of this cell type relative to the rest of the dataset, $q < 0.01$). Additional enriched TF motifs in cell type–specific peaks are provided in data file S3. (**C**) Examples of an (Left) activating versus (Right) repressive TF whose expression levels are positively or negatively correlated with motif accessibility across cell types and tissues. Each point indicates a cell type from a specific tissue [color code as in (B); shape code above]. Motif enrichment corresponds to fold-change of the mean motif occurrence in peaks of this cell type relative to the rest of the dataset. Expression values for the TFs are from sci-RNA-seq data collected in matching cell types as described in (*16*) (natural log of CPM+1). Correlation coefficient (R) values are Pearson correlations. (**D**) Correlation of motif enrichment and expression can be used to predict the mode of action of unclassified TFs. (Left) TFs were automatically assigned to the category of activator, repressor, or unclear on the basis of their associated GO terms. Pearson correlation values of motif enrichment and TF expression were calculated across all cell types in all tissues and are shown by category for all 455 TFs for which we have both values. Most TFs show positive correlation values. Annotated repressors have lower median R values than those of activators, with many of the outliers being due to missing or misleading GO term annotations. (Right) A high absolute R value can serve to classify TFs with unknown mode of action. An example is NFATc3, a likely repressor based on this analysis. (**E**) Position weight matrices (PWMs) identified by de novo motif search for exemplary cell types with no strong enrichment in (B). De novo motif enrichment was performed with homer (*48*) in the 2000 most specific peaks for each cell type by using CpG-matched genomic sequences as background. The closest known motif and the score for the motif matching process are indicated below. Further details as well as PWMs for all cell types are provided in fig. S7.

A central question in developmental biology is which TFs are involved in generating and maintaining a diversity of cell types from an invariant genome. We sought to leverage these data to systematically assess which TF motifs are differentially accessible and thus nominate key regulators of cell fate specification and/or maintenance in the context of in vivo human development. Differential motif accessibility is not proof of TF binding, so further experimental validation will be needed to confirm the below observations.

As a first approach, we used a linear regression model to ask which TF motifs found in the accessible sites of each cell best explain its cell type affiliation. Initially treating each tissue independently, we identified the most highly enriched motifs and TFs from the JASPAR database for each of 124 cell type clusters across all tissues, which revealed both known and potentially previously unknown regulators (fig. S5). For example, in the placenta, the motif of SPI1/PU.1, an established regulator of myeloid lineage development (*29*), is highly enriched in peaks of myeloid cells; the motif of TWIST-1, which is required for the formation of stromal progenitors (*30*), is enriched in peaks of stromal cells; and the FOS::JUN motif is associated with chromatin accessibility in extravillous trophoblasts, a cell type in which the corresponding AP1 complex has been described to be specifically active (*31*, *32*).

An unannotated cluster within the placenta is enriched for GATA1::TAL1 motifs, which are established regulators of erythropoiesis (*33*). These cells cluster with erythroblasts from other tissues in the global UMAP (Fig. 3A and fig. S6A), and upon further inspection, key erythroid marker genes exhibited specific promoter accessibility (fig. S6B). In the NNLS-guided workflow, this cluster was not annotated because an erythroblast cluster was not detected in the

placenta in the sci-RNA-seq study [possibly because the placenta is one of the few tissues for which we have more cells with ATAC than RNA data (*16*)]. Thus, motif enrichment can assist in cell type annotation, if the key regulators of a cell type are known.

We repeated this regression analysis on the 54 main cell types observed across all tissues, after collapsing cell types appearing in multiple tissues (Fig. 3B and data file S3; descartes.brotmanbaty.org) (*15*). As expected, the top motifs remained consistent with the tissue-specific analyses as well as the literature—for example, SPI1/PU.1 in myeloid cells (*29*), CRX in retinal pigment and photoreceptor cells (*34*), MEF2B in cardiomyocytes and skeletal muscle cells (*35*), and SRF in endocardial and smooth muscle cells (*36*). Whereas most motifs are enriched in only one or two cell types, neuronal TF motifs (*37*–*39*) are enriched in multiple neuronal cell types (Fig. 3B, top left cluster). Another exception to the cell-type specificity of motifs is HNF1B, which is conventionally associated with kidney and pancreas development (*40*, *41*) and whose motif is enriched in 13 cell types that span a range of specialized epithelial and secretory roles (*34*).

POU2F1 (POU class 2 homeobox 1) is an example of a TF that has not previously been associated with a particular developmental branch but rather has been suggested to be an exception within the POU family—broadly expressed and controlling no specific trajectory (*42*). By contrast, we found that in developing human tissues, its motif is enriched in several neuronal cell types. Lending further support, *POU2F1* is more highly expressed in those same cell types (fig. S6C).

Extending on this observation, we sought to leverage an atlas of gene expression (*16*) to more generally ask whether TFs are differentially expressed in a pattern consistent with the differential accessibility of their motifs. For example, looking across all cell types annotated in the same tissue in both datasets, the expression of the myeloid pioneer factor SPI1/PU.1 is strongly positively correlated with the enrichment of its motif at accessible sites (Fig. 3C, left). This analysis also revealed TFs with a negative correlation between their expression and motif enrichment (table S2). Upon closer inspection, these TFs tended to be repressors. For example, GFI1B has been described to act as a repressor crucial to erythroblast and megakaryocyte development by recruiting histone deacetylase upon binding its motif and inducing closing of the chromatin, such as at the embryonic hemoglobin locus (*43*). Consistent with this, we observed its expression to be negatively correlated with its motif enrichment at accessible sites (Fig. 3C, right).

Categorizing TFs as "activators" or "repressors" from GO terms, we found that TF expression and motif accessibility tend to be positively correlated for annotated activators and negatively correlated for annotated repressors (Fig. 3D, left). Exceptions can largely be explained by missing or conflicting GO terms, whereas literature searches are consistent with the observed correlation. Accordingly, this kind of analysis provides a systematic approach for classifying TFs as activators or repressors. For example, NFATc3 is generally described as an activator (*44*), but our analysis points toward a repressive mode of action, especially in developing T cells, where it is highly expressed yet its motif is depleted in accessible sites (Fig. 3D, right). Apart from a general classification, we also gained insight into the cell-type contexts in which a TF might variably act as an activator or repressor. For example, TFs including FOXO3 have been proposed to act as activators in their unmodified state but as repressors when phosphorylated (*45*), which might explain its more ambiguous relationship between expression and accessibility (fig. S6D). We only classified TFs as repressors if their presence is linked to a reduction in accessible chromatin, yet there are also TFs that have been reported to repress transcription while maintaining an accessible state at their binding sites, such as REST (*46*, *47*). This group of repressors is not distinguished from activators by our analysis (fig. S6E) because this would require further linking to the transcriptional effect on target genes.

A limitation of the above-described linear regression strategy for associating cell types with TF motifs is that it relies on databases of known TF motifs. As a different approach, we calculated specificity scores for each accessible site (*13*), selected the 2000 most specific peaks for each cell type, and searched de novo for enriched motifs within this set compared with CpG-matched background genomic sequences (fig. S7 and data file S4) (*48*). In general, the top de novo motifs for individual cell types agree with the top known motifs identified with linear regression. Some cell types that did not have strong matches to known motifs by means of the regression strategy were nonetheless associated with de novo motifs (such as endothelial, stromal, and Schwann cells) (Fig. 3E, and fig. S7). For endothelial cells in particular, this result is discussed further below.

Cross-tissue analyses of blood cells and endothelial cells

The nature of this dataset creates an opportunity to investigate organ-specific differences in chromatin accessibility within broadly appearing cell types such as blood and endothelial cells. In our first pass of cell type annotations for blood cells, we were able to differentiate between myeloid cells, lymphoid cells, erythroblasts, megakaryocytes, and hematopoietic stem cells (Fig. 2B). Extracting and reclustering these blood lineages from all organs allowed us to additionally identify macrophages, B cells, natural killer (NK)/type 3 innate lymphoid (ILC3) cells, T cells, and dendritic cells, once again adopting an RNA-assisted annotation approach (analyzing similar cell types from multiple tissues necessitated an additional doublet cleaning step) (Fig. 4A). Macrophages could be further separated into groups associated with tissue of origin, as previously observed (*49*), as well as phagocytic macrophages. This latter group was identified mainly in the spleen, followed by the liver and the adrenal gland (fig. S8A). In contrast to the RNA, we did not detect a separate microglia cluster in cerebrum, likely because this is a very rare cell type (~0.25%) (*16*).

Of particular interest within the blood lineages are erythroblasts, owing to the spatiotemporal dynamics of erythropoiesis during fetal development. We initially detected this lineage in the liver, adrenal gland, heart, and placenta (Fig. 2B); our cross-tissue analysis additionally identified erythroblasts in the shallowly profiled spleen (where only megakaryocytes and myeloid cells were originally annotated). The ratio of erythroblasts within the blood lineages of a tissue is highest in the liver, which is in line with this organ being the primary site of erythropoiesis at this developmental stage, followed by the spleen and adrenal gland (fig. S8A) (*50*, *51*), phenocopying the trend in the RNA data described in (*16*).

Further investigating erythroblasts, we observed that regions proximal to both the adult β- and fetal γ-globin genes are accessible at this stage of development, whereas the embryonic ε-globin gene's promoter is inaccessible (fig. S8B). The erythroblast cluster could be further subdivided into five major Louvain clusters with differential chromatin accessibility, including a distinct erythroblast progenitor cluster (Fig. 4A and fig. S8A). Accessible sites in the erythroblast progenitor cluster as well as in the adjacent early erythroblast cluster (erythroblast_3) are enriched for GATA1::TAL1 as well as other GATA motifs (Fig. 4B). Comparison of expression levels of various GATA factors in erythroblast progenitors allows us to nominate GATA1/2 as the TFs likely responsible for this motif enrichment (fig. S8C). The other erythroblast clusters, corresponding to later stages of erythropoiesis, show motif enrichment for NFE2/NFE2L2 (erythroblast_1) and NFYB/KLF1 factors (erythroblast_2/4) but a marked absence of enrichment for GATA motif accessibility. A scRNA-seq study on the mouse hematopoietic system reported induction of *GATA2* early in erythropoiesis, with a subsequent decrease in *GATA2* yet stable *GATA1* expression (*52*). By contrast, a study of sorted bulk human in vitro cultured erythroid populations revealed a decrease in *GATA1* expression from progenitors to differentiated erythroblasts, as well as increased *KLF1* and *NFE-2* levels in

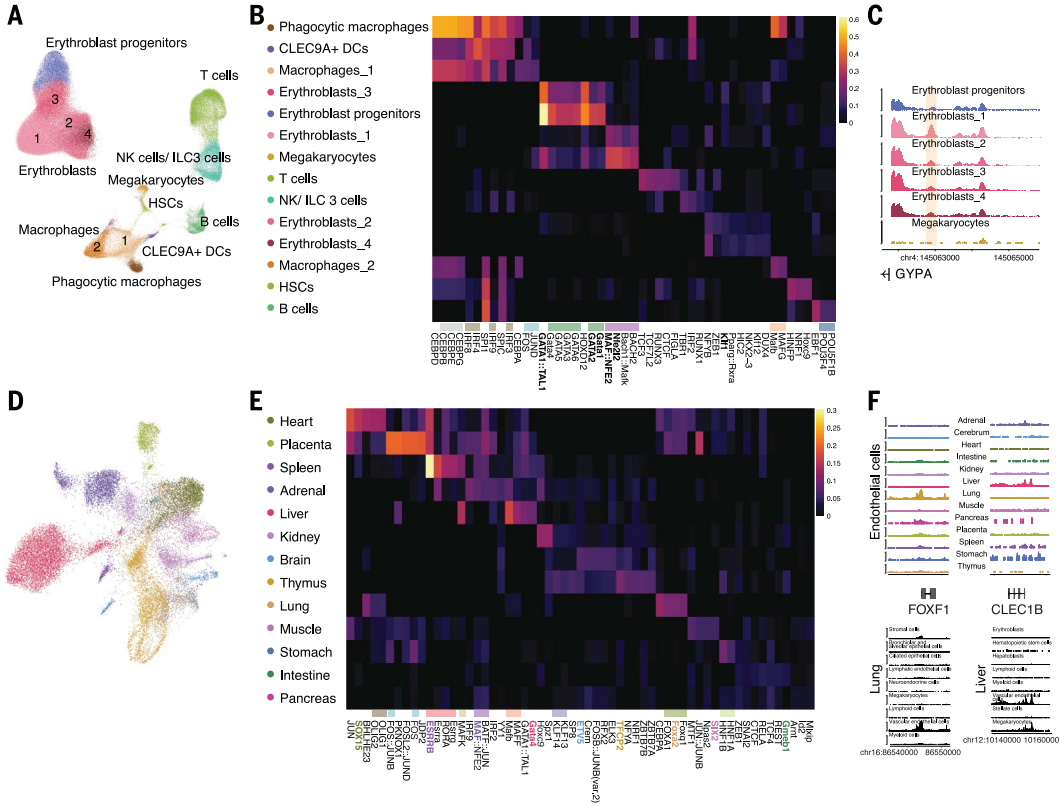


Fig. 4. Identifying major subgroups and associated TFs in broadly distributed lineages. (**A**) UMAP visualization of 152,649 blood cells extracted from all organs, colored and annotated by Louvain clusters. (**B**) Five TF motifs most strongly enriched in peaks of each Louvain cluster in (A) (log10-scaled fold-change of the mean motif occurrence in peaks of this cluster relative to the rest of the dataset, $q < 10^{-6}$). Highly similar motifs, as determined from RSAT matrix-clustering of the JASPAR vertebrate motif collection (*116*), are indicated with horizontal bars. (**C**) Example locus upstream of *GYPA* with differential accessibility across erythroblast populations. Accessibility is summed for all cells in each Louvain cluster, and the scale is normalized to account for differences in total reads per cell as well as cell numbers across clusters. Other blood cell types, including megakaryocytes (shown), have negligible accessibility at this region. (**D**) UMAP visualization of 27,576 vascular endothelial cells extracted from all organs and colored by tissue of origin. Colors are as in (E). Only the top 20,000 endothelial-specific peaks as determined in each tissue were used for clustering, merged to 94,023 distinct peaks across all tissues. (**E**) Five TF motifs most strongly enriched in peaks of each tissue group in (D) (log10-scaled fold-change of the mean motif occurrence in peaks of this tissue group relative to the rest of the dataset, $q < 10^{-4}$). Highly similar motifs, as determined from RSAT matrix-clustering of the JASPAR vertebrate motif collection (*116*), are indicated with horizontal bars. Motifs whose TFs (or TFs with highly similar motifs) are most highly expressed in endothelial cells from the same tissue in sci-RNA-seq data are highlighted (colors correspond to tissues). (**F**) Example loci showing specific accessibility in (left) lung or (right) liver endothelial cells. These sites also exhibit tissue-specific accessibility in their tissue of origin (bottom) and thus are unlikely to be consequent to residual doublets or free DNA contamination from other cell types. The *CLEC1B* locus is also accessible in the small cluster of megakaryocytes in liver and is known to be expressed in platelets (*117*). Accessibility is summed for all cells in each Louvain cluster, and the scale is normalized to account for differences in total reads per cell as well as cell numbers across clusters.

later-stage erythroblasts (*53*). Our observations align with the bulk in vitro human data on this point and indicate that there might be epigenetically distinct subpopulations of differentiated erythroblasts (subclusters 1, 2, and 4) in which the accessibility landscape is shaped by non-GATA factors (Fig. 4B). For example, a distal regulatory element upstream of *GYPA*, which is used as an erythrocyte invasion receptor by the malaria parasite (*54*), is most accessible in the erythroblast_1 population and contains a motif that resembles the NFE-2 motif (Fig. 4C).

Pseudotime analysis of hematopoietic stem cells (HSCs) and erythroblast subpopulations confirmed the order of progenitors and early erythroblasts in the HSC-to-erythroblast transition; late erythroblast clusters exhibited similar median pseudotimes, suggesting that they might represent subpopulations of differentiated erythroblasts rather than a succession of states (fig. S9, A and B) (*55*). This analysis also nominated candidate regulatory elements that open or close over the course of erythropoiesis (fig. S9C). Some of the top HSC- or erythroid progenitor–specific peaks (fig. S9C) are also accessible in bulk DNase profiles of fetal—but not adult—adrenal tissue (*56*–*58*), supporting the adrenal gland as a site of fetal hematopoiesis during normal mammalian development (fig. S9D) (*16*).

Another pervasive cell type is the vascular endothelium, which needs to perform both constitutive and highly specialized functions across organs, such as gas exchange in the lung or fluid filtration in the kidney. No TF has been described to be exclusively expressed in vascular endothelial cells, suggesting that the endothelial-specific transcriptome is controlled combinatorially by several TFs with overlapping expression in the endothelium (*59*). Consistent with this, we failed to observe any single, strong enrichment in endothelial cells in our analysis of JASPAR motifs (Fig. 3B). However, de novo motif discovery on the 2000 most endothelial-specific peaks revealed enrichment over background genomic sequences for motifs resembling ERG [E-26 transformation–specific (ETS)–related gene] and SOX15 [SRY (sex determining region Y)–box 15] (fig. S7). These motifs were likely not weighted as strongly in our linear regression approach because they are not restricted to endothelial cells (the ERG motif is enriched in megakaryocytes, and SOX15 is enriched in several cell types), nor is expression of these TFs limited to this cell type (fig. S10A). In line with this, ERG has been described as a major regulator of endothelial function (*60*) but also drives transdifferentiation into megakaryocytes (*60*, *61*).

We detected endothelial cells in 13 out of 15 organs, the exceptions being the more shallowly profiled cerebellum and eye (Fig. 2B). In contrast with erythroblasts (fig. S8A), extracting endothelial cells and reclustering revealed a marked separation according to tissue of origin (fig. S10B), in spite of stringent iterative filtering steps to remove residual contaminating doublets. Consistent with this, we also observed tissue-specific aspects of endothelial gene expression in fetal tissues (*16*) and previously

found regions exhibiting tissue-specific chromatin accessibility in adult mouse endothelial cells (*13*). To exclude technical sources for the tissue-specific signal, we selected the 20,000 most endothelial-specific peaks determined within each of the 13 tissues, merged these to 94,023 distinct peaks, and then clustered extracted endothelial cells on the basis of these peaks (Fig. 4D). The cells continued to cluster by tissue, similar to when we used all peaks (fig. S10B).

Further supporting tissue-specific differences in the endothelial regulatory landscape, endothelial cells derived from nearly all organs exhibited specific TF motif enrichments within these peaks (Fig. 4E). The TFs for many of the enriched motifs are also most highly expressed in endothelial cells of the matching tissue in the RNA data (Fig. 4E) (*16*). These analyses are limited to TF motifs present in the JASPAR vertebrate database, and additional TFs appear differentially expressed (*16*). Last, peaks of accessibility closest to differentially expressed genes have higher endothelial specificity scores in the matching tissue for about half of the profiled organs in the ATAC data (fig. S10C). Examples include *FOXF1*, which is specifically expressed and accessible in lung endothelium and whose promoter proximal region contains a FOXA2 motif; and *CLEC1B*, which is both specifically expressed in liver endothelium and harbors a GATA motif–containing candidate regulatory element exhibiting liver endothelium–specific accessibility (Fig. 4F). Some, but not all, of the enriched motifs are also enriched in other cell types of the same tissue. Although we cannot exclude residual contamination contributing to this signal, this might also reflect the underlying biology, for example, consequent to heterogeneous origins (*62*).

Overall, these findings indicate that the general program of chromatin accessibility and gene expression in endothelial cells, a widely distributed cell type that needs to fill both general and organ-specific functions, is mediated by a combination of constitutive TFs as well as tissue-specific TFs that may drive additional specialization. These analyses also highlight the merit of combining both de novo motif and linear regression approaches across tissues to nominate the key regulators that shape the chromatin landscape in individual cell types.

A catalog of accessible elements in the human genome during development

Altogether, our master set of 1.05 million sites spans 532 Mb, or 17.1% of the reference human genome (data file S1). This extensive catalog of accessible sites enabled several additional analyses. First, we used Cicero to generate coaccessibility and gene activity scores (*63*), analyzing each of 54 cell types separately. Because some of these were represented in several tissues, 101 Cicero maps were generated altogether. In total, we tested 159 million distinct pairs of accessible sites within 500 kb. At a coaccessibility score threshold (*63*) of 0.1, we obtained a catalog of 6.3 million distinct coaccessible pairs of sites across the 101 maps, with an average of ~139,000 pairs per cell type. This catalog includes 1.4 million (22%) promoter-distal, 4.8 million (76%) distal-distal, and ~94,000 (1.5%) promoter-promoter candidate interactions (data files S5 and S6; descartes.brotmanbaty.org) (*15*). For example, as expected at this stage of development, erythroblasts, but not other cell types, exhibited coaccessibility between the locus control region (LCR) and the fetal and adult, but not the embryonic, β-globin gene (fig. S11A) (*64*). A second example is the *FOXF1* promoter (Fig. 4F), at which endothelial cells from the lung, but not other tissues, exhibited coaccessibility with nearby distal elements (fig. S11B).

Second, a substantial proportion of heritability for common human diseases and traits partitions to accessible chromatin, particularly to regions that are specifically accessible in tissues or cell types related to the trait or disease in question (*65*–*67*). We previously intersected genome-wide association study (GWAS) signals for diverse human phenotypes with an adult mouse single-cell atlas of chromatin accessibility and found many anticipated relationships to be discoverable despite the considerable species difference (*13*). We repeated such an analysis on these data, applying partitioned linkage disequilibrium score regression (LDSC) (*67*) to detect enrichment of human heritability for 34 phenotypes from the UK Biobank (UKBB) within accessible chromatin for each of our 54 fetal cell types (Fig. 5A and table S3). Of the 54 cell types, 45 had a significant enrichment for at least one phenotype, whereas 32 of 34 phenotypes were enriched for at least one cell type (the exceptions being basal metabolic rate and sunburn, the latter in line with absence of skin tissue). As expected, for example, blood cell traits are maximally enriched in blood cell types, neurological phenotypes in neuronal cell types, and high cholesterol in hepatoblasts and intestinal epithelial cells. Further, type 2 diabetes is not only enriched in islet endocrine cells but also in pancreatic acinar and ductal cells, hepatoblasts, and stomach goblet cells; menopause age is maximally enriched in adrenocortical cells (fig. S11C). As similar single-cell atlases of chromatin accessibility are generated across the human life span, it will be interesting to explore at what time points these enrichments are maximal for each phenotype.

Third, we sought to evaluate the suitability of these data for identifying genetic variants that affect chromatin accessibility in cis. Although we generated data on many cells and tissues, they were collected from a relatively limited number of individuals, precluding the possibility of using an association framework. Instead, we sought to identify allelic imbalance within individuals at heterozygous positions (*68*). Specifically, we tested the liver and brain sample from two individuals, aggregating the reads for all cells from each cell type and testing for allelic imbalance across these aggregate measures. Overall, we found 586 single-nucleotide polymorphisms (SNPs) that exhibited a significant allelic imbalance [20% false discovery rate (FDR)] (tables S4 and S5). In general, the number of significant sites identified correlated with the number of reads from that cell type (fig. S12, A and B), and consequently, there were large differences in the power to detect allelic imbalance across cell types (fig. S12, C to F). Of the SNPs that were heterozygous in both individuals, sites that had significant imbalance in one individual were strongly enriched for significant imbalance in the same tissue in the other individual (49-fold over random for brain, hypergeometric test $P = 1.5 \times 10^{-36}$; 59-fold over random for liver, hypergeometric test $P = 2.3 \times 10^{-60}$), although there was a greater degree of sharing between the liver and brain of the same individual (69-fold over random for one individual, hypergeometric test $P = 1.2 \times 10^{-77}$; 78-fold enrichment, hypergeometric test $P = 5.7 \times 10^{-44}$ for the other individual). Although not significantly enriched ($P = 0.059$), 25 SNPs with allelic imbalance in at least one cell type were previously associated with complex traits in the National Human Genome Research Institute–European Bioinformatics Institute (NHGRI-EBI) GWAS catalog (table S6) (*69*). For example, rs61851769 shows allelic imbalance in erythroblasts and hepatoblasts in one liver sample and was previously associated with mean corpuscular hemoglobin (Fig. 5B) (*69*, *70*). The SNP disrupts a TAL1 binding site and is upstream of *SLC30A1*, a gene implicated in erythropoiesis (*71*). Consistent with the erythroblast-specific nature of these annotations, we believe that the hepatoblast signal may come from contaminating erythroblasts because hepatoblast accessibility is lost after peak module–based doublet filtering. Another example is rs362649, which is significant in excitatory neurons of one individual, was previously associated with the volume of cerebellar vermal lobules VIII to X (*72*) and lies within an intron of *RELN*, which plays a role in neuronal migration (Fig. 5B) (*73*). There are many caveats to these analyses, including the large differences in power across cell types. Nonetheless, these results illustrate how single-cell chromatin accessibility data might be leveraged for the identification of functional noncoding genetic variation with cell-type resolution.

Fourth, analogous to grouping cells on the basis of their shared patterns of accessibility across sites (Fig. 3A), we can instead group sites by their shared accessibility across cells (*74*, *75*). To reduce the computational complexity of this task, we removed sites of <400 base pair (bp) width and then computed a "UMAP of

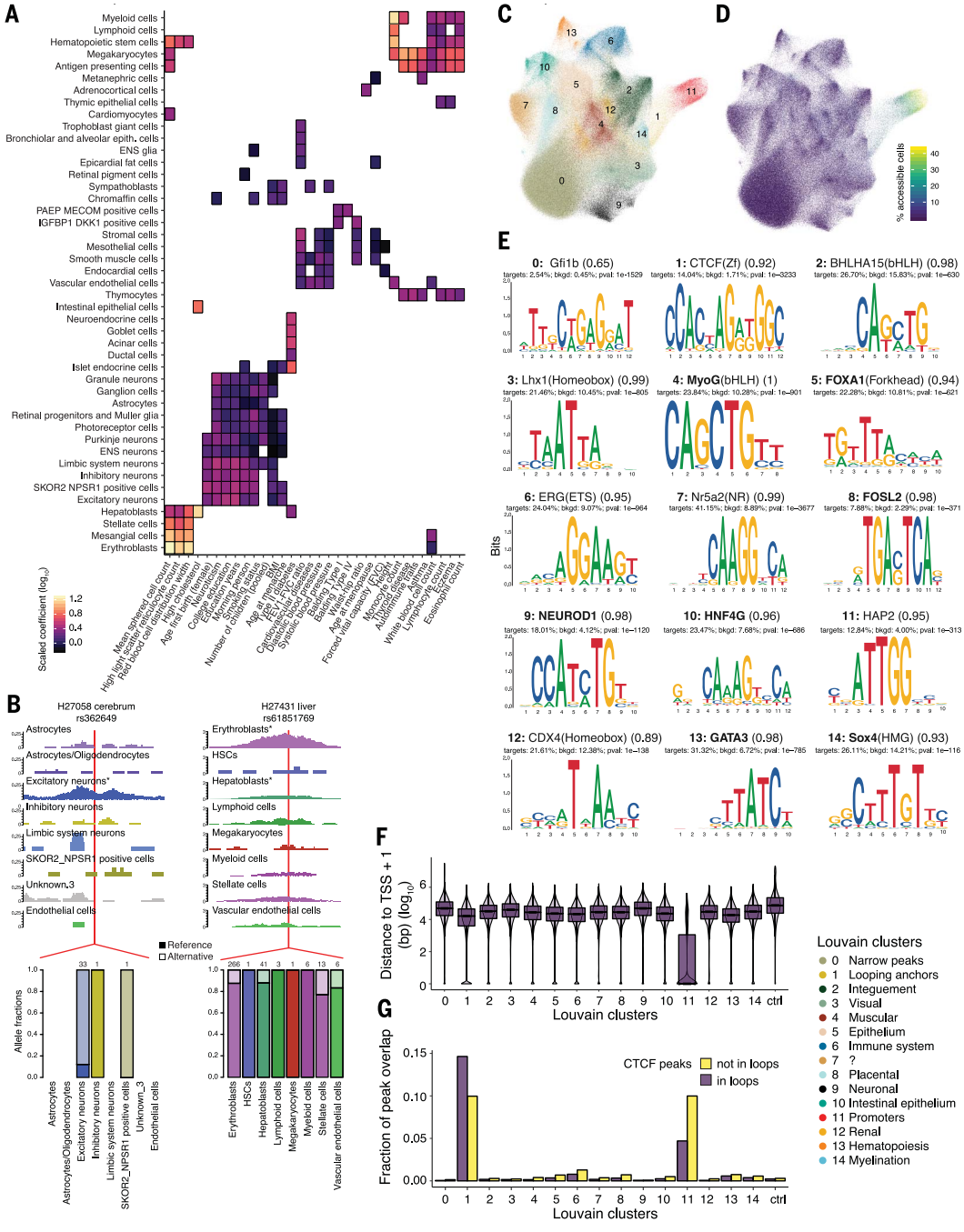


Fig. 5. Heritability enrichment and coaccessibility of candidate regulatory regions. (**A**) Enrichment of heritability for UK Biobank traits within top 10,000 specific sites for each cell type. Trait–cell type pairs with no significant positive enrichment ($q > 0.2$) are white. A full table of scaled coefficients and q values for each trait–cell type pair is provided in table S3. (**B**) Example sites with allelic imbalance. Browser tracks of accessibility for the cell types in a (left) cerebrum and (right) liver sample are normalized to counts per million reads. Results are presented as unsmoothed base coverage. Asterisks indicate cell types with significant allelic imbalance. The red vertical line indicates the position of the SNP exhibiting allelic imbalance. The bar plots below show the relative portions of reads mapping to the reference and alternative allele at that position. Above each bar is the number of reads overlapping the SNP for each cell type. (**C**) UMAP visualization of a subset of accessible regions from the master set that are >400 bp (447,879 sites), by using accessibility profiles from the subsampled cell dataset in Fig. 3B (88,983 cells). Sites are colored by Louvain clusters, which are numbered according to decreasing size and annotated into broad categories on the basis of motif enrichment and lineage affiliation of enriched cells. Legend is at bottom right of the overall figure. Cluster 0 consists of narrower sites with the lowest accessibility across cells, is not enriched for a clear motif, and possibly reflects rare or transient cell states or biological or technical noise. (**D**) Same as (C), but colored by the percentage of cells in which sites are accessible. A version in which the accessible percentage is binned by content is shown in fig. S13C. (**E**) PWMs identified by means of de novo motif search in each of the clusters in (C). De novo motif search was performed with homer (*48*), using CpG-matched genomic sequences as background. The top PWM per cluster 0 to 14 is labeled by the closest known motif as determined by homer, with the score for the motif matching process indicated in brackets. Listed below are the percentage of sites within the cluster and CpG-matched background sequences that contain a match to the de novo PWM, and a P value for the enrichment. Motifs associated with pioneer TFs are in boldface. The top motif for cluster 0 is only found in 2.5% of sites and has a poor matching score. (**F**) Violin plots of the distances of each group of sites in (C) to the nearest TSS is shown. 20,000 random regions located on autosomes with a width corresponding to the median width of all sites in (C) were used as control (ctrl). (**G**) Fraction of sites within each cluster overlapping with ENCODE-defined CTCF-bound peaks within versus outside of looping regions. All CTCF ChIP-seq peaks overlapping CTCF motifs in looping regions in GM12878 (n = 8253 peaks) and the same number of ChIP-seq peaks not overlapping looping regions but with the same ChIP-seq score were selected. For each cluster in (C), the fraction of sites overlapping these two CTCF-bound sets was calculated. The same control (ctrl) regions as in (F) were used.

sites," grouping 447,879 regions into 15 clusters (Fig. 5, C and D). Applying the aforedescribed linear regression and de novo motif search strategies, most of these 15 clusters were enriched for key TF regulators identified by our earlier analyses (Fig. 5E, fig. S13A, and data file S7). Correspondingly, when we determined "differential cells" (analogous to determining differential genes or peaks in conventional clustering of single-cell data), we found that cells from lineages that match the motif enrichments define most of these clusters (fig. S13B). Thus, most of these clusters represent sites specifically accessible in certain cell types or cell-type groups and therefore link to cell type–defining TFs.

The top cluster-defining TFs identified through de novo motif search include several pioneer factors, implying that sites bound by these TFs are more likely to be concurrently accessible.

However, a few of the clusters of sites were not enriched in a pattern that reflected a specific lineage. For example, cluster 11, comprising 10,983 or 2.5% of sites, clearly corresponds to commonly accessible promoters: Its sites are accessible in many cells (Fig. 5D and fig. S13C); 75% are within 1 kb of a TSS (Fig. 5F); and they are broader, CpG rich, and conserved (fig. S13D). In addition, this cluster is strongly enriched for motifs commonly found in promoters—such as various SP factors, KLF factors, NRF1, and ZFX (fig. S13A)—and the top identified de novo motif corresponds to the CCAAT promoter element (Fig. 5E). In particular, this cluster is enriched for housekeeping gene promoters [1.9-fold enriched, hypergeometric test $P = 6.5 \times 10^{-244}$; 80% of 3006 housekeeping TSSs defined by (*76*) are in this cluster].

Another case is cluster 1, whose 41,128 sites are not as commonly accessible as those of promoters (Fig. 5D) but are nonetheless less cell type–specific than other clusters (fig. S13B). These sites also have higher CpG content and are modestly broader and slightly nearer to TSSs than other nonpromoter clusters (Fig. 5F and fig. S13D). Although this might reflect a cluster of sites containing some promoters, motifs of promoter TFs are depleted in cluster 1 (data file S7). Its only significantly enriched motif is CTCF (Fig. 5E and fig. S13A). This suggests that these coaccessible sites correspond to TAD (topologically associating domain) boundaries and looping anchors, which are known to bind CTCF and to be largely but not entirely invariant across cell types (*77*).

To evaluate this hypothesis, we obtained CTCF-bound peak locations from ENCODE, as determined with chromatin immunoprecipitation–sequencing (ChIP-seq), as well as loop anchor locations from Hi-C data in GM12878 (*78*), and calculated the overlap of each cluster of sites with CTCF-bound peaks within versus outside of looping anchors (Fig. 5G). Most clusters showed limited overlap. A first exception was cluster 11 (promoters; 10% overlap with nonlooping peaks), which is in line with 20% of CTCF sites falling in promoters (*79*). A second exception was the CTCF-enriched cluster 1 (15% overlap with looping peaks, a number that would likely increase if Hi-C and ChIP-seq data from all profiled cell types were available). This was also the only cluster with greater overlap with looping than nonlooping CTCF-bound peaks. Taken together, profiling chromatin accessibility across many tissues reveals not only cell types but also sets of coaccessible regulatory elements—mostly lineage-specific sets, but also promoters and looping regions.

Fifth, we compared our master list of sites to orthogonally annotated functional regulatory regions in the human genome and accessible regions in other species. Of human accelerated regions (*80*), 66% overlap one of our peaks, as do 75% of human VISTA enhancers (*81*). Nonoverlapping VISTA enhancers are slightly enriched for an absence of expression in transgenic mouse assays (1.2-fold; hypergeometric test $P = 6.9 \times 10^{-8}$). Peaks that we assigned to the visual, neuronal, and looping categories (Fig. 5C) are enriched for overlap with both human VISTA enhancers and accelerated regions, whereas narrow, rarely accessible peaks are depleted (fig. S14, A and B). We also compared our master list of sites to the peak set generated by profiling chromatin accessibility in 13 tissues from 8-week-old mice (*13*). Of the 23% of these mouse peaks that lift over to the human genome, 60% (61,396) overlap a human peak. The overlapping human peaks are significantly enriched for peaks associated with neuronal or myelination cell types, looping anchors, and promoters but not other cell types (such as immune or hematopoiesis); narrow rare peaks are depleted, as are placental peaks (placenta was not profiled in the mouse atlas) (fig. S14C). The result is consistent with the possibility that regulatory sites of some broad categories of cell types (such as neuronal cells) may have experienced less evolutionary turnover between mouse and human than others (such as immune cells) (*17*, *82*).

Comparisons of accessibility across developmental stages

We next asked whether cell type–specific motif enrichments are shared across developmental stages. Many similar cell types show similar top motifs enriched in the mouse ATAC atlas, which was generated by profiling 13 tissues in 8-week-old mice (*13*), implying that these TFs have a role in cell fate maintenance that may be conserved across species (mouse versus human) as well as developmental stage (adult versus fetal) (fig. S15, A and B). POU2F1—the motif we suggest to be important for neuronal cells—is enriched in accessible sites of mouse excitatory neurons, in addition to B cells (fig. S15A). Motif enrichment patterns cluster largely by cell type rather than species in a shared heatmap (fig. S15C), with some exceptions. For example, whereas myeloid cells cluster together, mouse lymphoid cells cluster separately from human lymphoid cells, in part driven by a more pronounced enrichment of NFKB1/2 motifs in the mouse. This could be due to the difference in developmental stage because NFKB1 has been shown to be dispensable for the emergence of prenatal B-1 transitional cells yet essential later in development (*83*).

To investigate human developmental stage–specific chromatin accessibility, we compared our dataset with existing single-cell ATAC datasets in adult human tissues, namely blood and cortex (*84*, *85*). To remove strong batch effects observed, we selected overlapping peaks in the adult dataset, rescored our data on the basis of this peak set, identified anchors, and integrated the two datasets (fig. S16A) (*23*). After applying this integration strategy, blood cells clustered by cell type rather than stage, with fetal cells falling closer to naive subtypes in the UMAP visualization (fig. S16B). As with the comparison with the mouse atlas, and as expected given the relatively late stage of development that we were interrogating, we observed similar motifs enriched in many blood cell types, with some differences (fig. S16C). Again, adult B and T cells are more strongly enriched for NFKB1/2 (1.5- to 1.6-fold for adult B and T cells and 1.1-fold for fetal B cells; fetal T cells showed no enrichment). However, such comparisons are hampered by strong batch effects owing to different sample collection and processing as well as the removal of potentially meaningful dataset-specific differences in the integration workflow.

In our comparison of the developing versus adult cortex data (*85*), again several cell types overlap in the integrated UMAP representation (fig. S17A). However, some fetal subgroups, including the two largest excitatory neuron subgroups, do not overlap with the adult data (subgroups 1 and 2) (fig. S17B). The fetal cerebral UMAP contains more substructure than we annotated (as do other tissues and cell types), evidenced by cluster-specific accessibility at known neuronal subtype marker genes (Fig. 2B and fig. S17C). For a more in-depth analysis of one of the cell types, we sought to further annotate subgroups of the most prevalent fetal cerebral cell type: excitatory neurons. To this end, we first applied our NNLS-based cell-type annotation strategy using single-cell expression data from the Allen Brain Atlas, which was collected on post mortem adult brain samples (*86*). Whereas many clusters found a match, the largest excitatory neuron subgroup did not (subgroup 1) (fig. S17D). By contrast, when using single-cell expression data collected from developing cortex (gestational week 17 to 18) (*85*, *87*), we found that the two largest excitatory neuron subclusters match to newly formed migrating and maturing excitatory neurons, respectively (subgroups 1 and 2) (fig. S17E). Of the top 10 peaks specific to the migrating population (subgroup 1), four lie in the introns of neuronal genes, four lie in the introns of noncoding RNAs, and two are distal to transcriptional units but highly conserved in vertebrates (fig. S18, A and B). One example of the latter is a distal peak (>20 kb from nearest TSS) overlapping a conserved element listed as negative in the VISTA enhancer browser (fig. S18A) (*80*).

This finer annotation also enabled us to ask whether heritability for certain traits is differentially enriched across neuronal subtypes. We calculated enrichments of trait heritability for each Louvain cluster in the cerebrum,

instead of each cell type, compared with the entire dataset. As expected, we observed enrichment for various neurological traits in the neuronal cell types but not in non-neuronal cell types, such as brain endothelial cells (fig. S19A). Within our broadly annotated cell types, we observed variable enrichment for different Louvain clusters; for example, inhibitory neuron subtype 2 is strongly enriched for heritability of both bipolar disorder and number of children born to males (fig. S19B). As for the excitatory neurons, we found that heritability for educational attainment is more strongly enriched in accessible sites of differentiated deep-layer excitatory neurons than migrating or maturing excitatory neurons (fig. S19C). Conversely, anorexia heritability is more strongly enriched in accessible sites of maturing excitatory neurons (fig. S19C).

An inspection of TF motifs differentially enriched across excitatory neuron subgroups revealed that POU2F1, which we showed may be restricted to neurons, is most strongly enriched in the fetal-specific migrating group, suggesting that it might not only be involved in maintenance but also specification of neuronal fates (Fig. 6A). In line with this, an enhancer adjacent to *POU2F1* has been shown to be specifically active in mouse cortical progenitor cells (*88*). To further investigate the regulatory landscape during excitatory neuron development, we next generated a pseudotime trajectory, from migrating over maturing to differentiated deep-layer neurons (Fig. 6, B and C). Differences in the median pseudotime of all excitatory neuron cells from individual donors corresponded loosely to differences in gestational age (Fig. 6C), although the number of individuals ($n = 3$) was too small for detailed investigation of this interindividual heterogeneity. Thousands of excitatory neuron peaks open or close in a pseudotime-dependent manner (Fig. 6D). Dynamically accessible elements that open over pseudotime were enriched for motifs of Rfx- and Tal-related factors important for neuronal maturation, whereas elements that close over pseudotime are enriched, among others, for motifs belonging to paired-related homeodomain factors and POU factors, including POU2F1 (2.2-fold change, $q = 2.3 \times 10^{-4}$) (Fig. 6E). This dynamic is supported by the matched RNA data, in which *POU2F1* expression peaks early in the pseudotime trajectory of excitatory neuron development (Fig. 6F) (*16*). Similar analyses of developmental-specific cell populations, their associated candidate regulatory regions, and TF motifs could be applied to further tissues once the progenitor population has been identified.

Discussion

Sci-ATAC-seq3 adds to a growing repertoire of single-cell methods that use combinatorial indexing, a technical paradigm whose advantages over other platforms include exponential scaling and greater range with which to profile diverse aspects of single-cell biology (*1*–*12*). Although libraries have limited complexity and sci- protocols suffer from loss of material during the pooling and washing steps, the results presented here and in (*16*) illustrate the power of sci- methods. All experiments were conducted by a handful of individuals in a nonproduction environment but nonetheless resulted in very large single-cell chromatin accessibility and gene expression datasets.

An overarching goal of the field is to develop an "atlas" of human gene regulation as it unfolds across development and across the human life span. Aside from scale, our studies differ from other recent single-cell atlasing reports in at least three respects. First, we sought to profile as many tissues as possible within the context of a single study rather than focus on a single organ. This was both to create a broadly useful reference atlas as well as to enable cross-tissue comparisons of widely distributed cell types. For example, we observed tissue-specific chromatin accessibility and gene expression for endothelial cells but not erythroblasts.

Second, we focused on tissues obtained during human development. The rationale for this choice is discussed in greater length in (*16*) but includes our goal of laying a foundation for the systematic investigation of genetic disorders of development, which account for a disproportionate proportion of pediatric disease (*89*, *90*). The further accumulation of similar data from additional developmental time

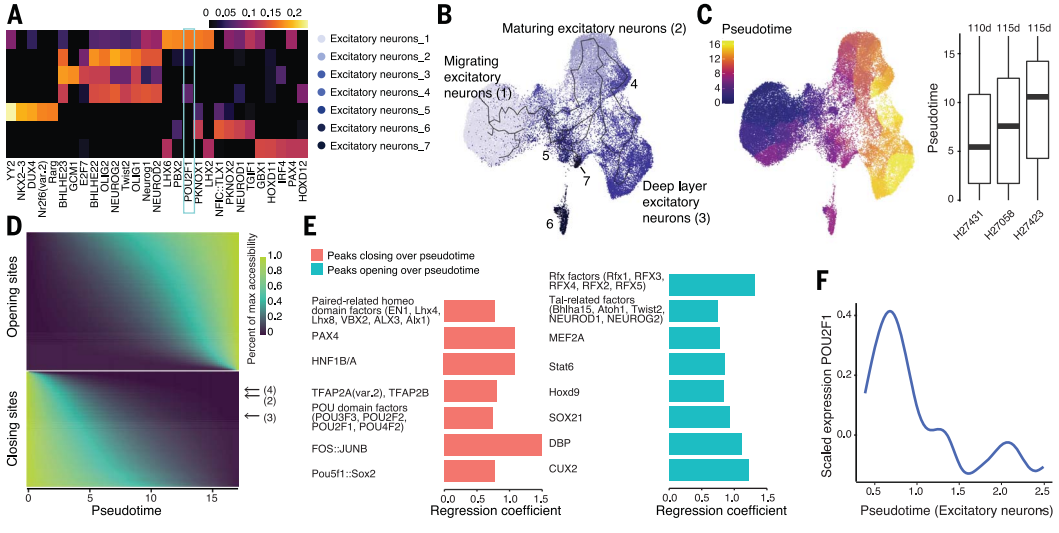


Fig. 6. Chromatin accessibility dynamics in developing excitatory neurons. (**A**) TF motifs enriched in excitatory neuron clusters. Fold-change of the top five enriched TF motifs in cluster-specific peaks for each of seven Louvain clusters that were annotated as excitatory neurons (log10-scaled fold-change of the mean motif occurrence in peaks of this cell subtype relative to the rest of the excitatory neurons, $q < 0.01$). POU2F1 enrichment is highlighted with a vertical box. (**B**) UMAP visualization and pseudotime trajectory path of 48,733 excitatory neurons colored by Louvain cluster. Color legend is in (A). (**C**) Pseudotime of excitatory neurons. (Left) UMAP visualization colored by pseudotime and (right) boxplots of median pseudotime per individual donor. Estimated gestational age is indicated above the boxplots. (**D**) Smoothed pseudotime-dependent accessibility curves of excitatory neurons, generated by a negative binomial regression and scaled as a percent of the maximum accessibility of each site. Sites (rows) are sorted by the pseudotime at which they first reach half their maximum accessibility. A random 10% of accessible sites was selected, and 3387 sites with pseudotime-dependent accessibility ($P < 0.05$, Wald test) are shown. Peaks from fig. S18A are indicated with arrows. (**E**) Motif enrichments in dynamically accessible sites from (D). Coefficients from logistic regression model by using the presence or absence of a given motif in each site to predict whether the site has a given accessibility trend. Plots show the top motifs ordered by Benjamini-Hochberg corrected q value for each category ($q < 0.05$); similar motifs are grouped together. (**F**) Expression dynamics of *POU2F1* over pseudotime in excitatory neurons. Smoothed *POU2F1* expression in matching excitatory neurons from (*16*) was normalized by size factor in each single cell, then log-transformed and scaled.

points in both mouse and human will enable a systematic understanding of in vivo emergence and differentiation of mammalian cell types.

Third, we chose to study not only single-cell gene expression but also chromatin accessibility, in the same tissues and where possible from identical samples (*16*). Genomic regions exhibiting cell type–specific chromatin accessibility generally correspond to DNA regulatory elements such as enhancers and thus afford the opportunity to understand not only the "output" of the genome in particular cell types but also the regulatory program that underpins that output. The aggregate of all accessible regions that we identified spans 17% of the human genome, which is in line with recent bulk DNase-seq profiles from fetal tissues (*91*). Most of these ~1 million elements are cell type–specific or cell type–restricted in accessibility, although a large group of shared elements likely corresponds to looping anchors. Further studies (for example, those based on evolutionary conservation, massively parallel reporter assays, and/or CRISPR perturbation) are necessary to validate these candidate regulatory elements as well as their Cicero-based candidate linkages to target genes.

An interactive website facilitates the exploration of these data by tissue, cell type, locus, or motif (descartes.brotmanbaty.org) (*15*). Beyond constituting a rich and easily accessible resource for the field (for example, providing individual researchers with information on their gene, enhancer, or cell type of interest), this dataset also enables us to learn about more general aspects of gene regulation. For example, leveraging that we have matching chromatin accessibility and gene expression data spanning so many tissues and cell types allows us to study the mode of action of TFs as well as organ-specific differences in the regulatory landscape of cell types or cell type–specific disease heritability. Because the underlying methods are relatively new, there is currently a paucity of single-cell chromatin accessibility datasets in the public domain. We anticipate further comparisons to adult humans (*92*) or other species (*13*) as more such data become available.

The breadth and resolution of this dataset also provides insights into specific developmental processes. POU2F1 is one of the earliest described mammalian TFs (*93*). It is thought to be the only known POU family member not expressed in a specific temporal or spatial pattern, and in spite of being the subject of many studies, to date POU2F1's role has remained elusive (*42*, *94*). Although it has been suggested to be involved in housekeeping gene regulation or tumorigenesis, knockdowns in cancer cell lines showed no growth defect (*42*). The single-cell resolution provided by this study reveals that *POU2F1* is more highly expressed in neuronal cell types, and its motif is specifically enriched in neuronal regulatory regions. Because we captured developing neurons in our profiling window, we could observe that this motif is most highly enriched in the developing population of excitatory neurons, which is mirrored by *POU2F1* expression dynamics. *POU2F1* and its binding sites are highly conserved (*42*), and we also observed motif enrichment in mouse excitatory neurons, implying that this TF is a conserved inducer and maintainer of excitatory neuron cell fate. In line with this, POU2F1 deficiency is embryonic lethal (*95*). This example illustrates the power of combined chromatin accessibility and gene expression data at single-cell resolution. We anticipate that further such examples will emerge with more in-depth analyses of other tissue systems, stages, and cell types.

These and other downstream analyses used stratifications of accessibility that were based on our cell-type annotations. Although our assignments appear appropriate given that they generally recapitulate known biology in downstream analyses, they should be regarded as preliminary and will likely necessitate adjustments as more atlases and improved data become available. We intentionally kept our cell-type annotations rather broad, but there is more substructure in the data that could be explored further by subclustering—for example, as we show for blood cells and excitatory neurons. Although we are undoubtedly missing many cell types because of shallow profiling of several tissues or insufficiently aggressive clustering, we were nonetheless able to derive chromatin accessibility profiles and key regulators for some rare and potentially previously unknown cell types.

The analyses that we present here are only a starting point. Many other facets can be explored directly from these data—for example, nominating sets of TFs that must be coexpressed in the same cell type in order to bind regulatory regions cooperatively. In addition, these data can directly be used as input to machine learning models—for example, to predict the effect of all disease-associated variants identified in the human genome on chromatin accessibility across all cell types (*96*). We foresee that the true power of single-cell methods will lie in combining descriptive resources as we present here with both machine learning and high-throughput perturbation, with the long-term goal of establishing a deep, predictive understanding of gene regulation in human development and disease.

Materials and methods

A more detailed version of materials and methods is provided as supplementary materials.

sci-ATAC-seq3

A more detailed version of the full sci-ATAC-seq3 workflow is available on protocols.io (*97*) and in the supplementary materials.

Preparation of nuclei

Human fetal tissues (89 to 125 days estimated post-conceptual age) were obtained by the University of Washington Birth Defects Research Laboratory (BDRL) under a protocol approved by the University of Washington Institutional Review Board. Tissues of interest were isolated and rinsed in 1X Hanks' balanced salt solution. Dried tissue was snap frozen in liquid nitrogen, manually pulverized on dry ice with a chilled hammer, aliquoted, and stored at −80°C until further processing. A subset of these aliquots were used for sci-ATAC-seq3, and others were used for sci-RNA-seq3, as described in the companion paper (*16*). For ATAC, nuclei were lysed with Omni-ATAC lysis buffer (*98*), cross-linked with 1% formaldehyde, and snap frozen in freezing buffer (*99*).

sci-ATAC-seq3 library construction and sequencing

Frozen fixed nuclei were thawed, resuspended in Omni lysis buffer (*98*), and diluted in ATAC–resuspension buffer (RSB) buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MaCl2) supplemented with 0.1% Tween-20. For three-level indexing experiments at 384^3, the nuclei input number was 4.8 million at 50,000 nuclei per well spread across 96 reactions. We profiled 24 individual tissue samples per batch (table S1), in which the 24th sample was a mixture of sentinel tissue (trisomy 18 cerebrum) and a mouse cell line (CH12-LX). For each sample, 200,000 nuclei were pelleted and resuspended in tagmentation reaction master mix (Nextera TD buffer, 1X Dulbecco's phosphate-buffered saline, 0.01% Digitonin, 0.1% Tween-20). Nuclei in tagmentation reaction master mix were aliquoted into four wells per tissue sample across a LoBind 96-well plate, 2.5 μl of Nextera v2 enzyme were added per well, and the plate was incubated at 55°C for 30 min. Tagmentation reactions were stopped by adding stop reaction mixture (40 mM EDTA with 1 mM Spermidine) and incubating at 37°C for 15 min. Tagmented nuclei from each sample were pooled (24 sample tubes in a batch), pelleted, washed, and resuspended in ATAC-RSB with 0.1% Tween-20. After adding phosphorylation master mix [1X polynucleotide kinase (PNK) buffer, 1 mM rATP, T4 PNK], the phosphorylation and nuclei reaction mix was aliquoted across a total of 16 wells in four LoBind 96-well plates and incubated at 37°C for 30 min. Ligation master mix (1X T7 ligase buffer, N5_splint, T7 DNA ligase enzyme) was added to the nuclei in the phosphorylation reaction followed by N5_oligos (384 distinct N5 barcodes). Sequences of all splint and barcode oligos used for sci-ATAC-seq3 are provided in table S7. Plates were incubated at 25°C for 1 hour. After this first round of ligation, stop reaction mixture was added, and the plates were incubated at 37°C for 15 min. Wells

were pooled, and nuclei were transferred into a 50-ml falcon tube, pelleted, and washed with ATAC-RSB with 0.1% Tween-20. The nuclei were then resuspended in N7 ligation master mix (1X T7 ligase buffer, N7_splint, T7 DNA ligase). This ligation and nuclei master mix was aliquoted into four 96-well LoBind plates, and N7_oligos (384 distinct N7 barcodes) were added to each well across four 96-well plates. Plates were incubated at 25°C for 1 hour before adding stop reaction mixture and incubating the plates at 37°C for 15 min. Wells were pooled and nuclei transferred into a 50-ml falcon tube, pelleted, and resuspended in Qiagen EB buffer. Then, 1000 to 3000 nuclei were aliquoted per well across four 96-well LoBind plates. To reverse cross-link the nuclei, we added a reverse cross-link master mix of EB buffer, PNK, and 1% SDS to each well. Plates were incubated at 65°C for 16 hours. A test PCR amplification was performed, and the reaction was monitored with SYBR green on several wells of a plate to determine the optimal cycle number (*1*). On the basis of this test PCR result, the rest of the reversed cross-linked plates were amplified with Nextera PCR Mastermix (NPM), bovine serum albumin, indexed P5 oligo, and indexed P7 oligo. Amplification products were pooled and purified first by using Zymo Clean & Concentrate-5 and then 1X AMPure beads. Final libraries were quantified on an Agilent 4200 Tapestation System. A 2 nM pool was created from equimolar pooling and sequenced with custom recipe and primers (sequences are provided in table S7) on an Illumina NovaSeq 6000 sequencer with custom sequencing recipe (read 1: 51 cycles, read 2: 51 cycles, index 1: 10 cycles+15 dark cycles+10 cycles, index 2: 10 cycles+15 dark cycles+10 cycles).

Data processing for sci-ATAC-seq3

A more detailed version of all data processing and analysis steps is available in the supplementary materials. A demultiplexing script and tutorial are provided on Zenodo at (*100*).

Data processing for the barnyard experiments conducted to develop sci-ATAC-seq3 was done as previously described (*13*). Methods for processing sequencing data from the tissue samples closely follow the methods used in (*13*) as well, albeit with numerous optimizations to scale to larger datasets.

Cell barcodes were separated from the distribution of background barcodes by fitting a mixture of two negative binomials (noise versus signal). Nonduplicate fragment endpoints for each cell were used for peak calling in each sample by use of MACS2 (*101*). Peak calls from all samples included in downstream analysis were merged to form a master set of peaks. For each sample, we created sparse matrices counting (i) reads falling within the master set of peaks and (ii) reads falling within gene bodies extended by 2 kb upstream for each cell. We additionally tabulated the total number of reads from each cell coming from annotated TSSs (±1 kb around each TSS), ENCODE blacklist regions, and our set of merged peaks for quality control (QC) purposes. To filter out low-quality cells, we chose tissue-specific cut-offs for the fraction of nonduplicate reads in peaks (minimums ranging from 20 to 40%) and the fraction of nonduplicate reads falling in TSSs (minimums ranging from 5 to 15%) by means of visual inspection of their distributions for each sample (for example, for some tissues we observed a bimodal distribution for the fraction of nonduplicate reads in peaks and removed the lower mode), and a global cutoff of 0.5% of nonduplicate reads coming from ENCODE blacklist regions. All downstream steps were performed one tissue at a time by pooling cells passing QC from all samples of a given tissue. We used a modified version of the scrublet (*102*) algorithm to remove the cells most likely to be doublets.

QC of sci-ATAC-seq3 data in bulk

After initial processing of the data, we assessed its quality relative to bulk DNase-seq profiles generated on fetal tissues procured from BDRL by the Roadmap Epigenomics consortium (*58*). After reprocessing the DNase-seq data in a comparable manner, we generated a master list of peaks across all DNase-seq and sci-ATAC-seq3 samples by merging all peaks called on each individual sample and generated a matrix of reads by peaks for each sample. This matrix of read counts was then used to calculate pairwise Spearman correlations to evaluate how similar samples were in their distributions of accessibility.

Dimensionality reduction and clustering

For dimensionality reduction, we found that the implementation of latent semantic indexing [LSI; or equivalently, latent semantic analysis (LSA)] that we have previously applied (*13*) did not perform well on data collected in this study, likely owing to sparsity. Log-scaling the term-frequency term in LSI resulted in very similar performance to those of the other tools we tested (*103*, *104*). We suspect that this is due to the exponential distribution of total counts per cell and the impact of strong outliers on the principal components analysis (PCA) step of LSI in the absence of log scaling.

We performed LSI on the binarized peak-by-cell matrix for all cells passing QC from each tissue, one tissue at a time. We first weighted all the sites for individual cells by log(total number of peaks accessible in cell) (log-scaled "term frequency"). We then multiplied these weighted values by log(1 + the inverse frequency of each site across all cells), the "inverse document frequency." We used singular value decomposition on the term frequency–inverse document frequency matrix to generate a lower-dimensional representation of the data (PCA) by only retaining the 2nd through 50th dimensions (the first dimension tends to be highly correlated with read depth). L2 normalization was performed on the PCA matrix to further account for differences in the number of nonduplicate fragments per cell. This L2-normalized PCA matrix was used for all downstream steps.

Although we did not observe evidence for substantial batch effects between samples, we nonetheless applied the Harmony batch correction algorithm on the PCA space to correct batch effects between different samples (*20*, *21*). This corrected L2-normalized PCA space was used as input to Louvain clustering and UMAP as implemented in Seurat V3 (*105*).

Cell type annotation

To transfer cell type labels for our Louvain clusters from the companion sci-RNA-seq data, we used an NNLS-based cluster-by-cluster annotation approach, which we have implemented previously to transfer labels between single-cell RNA-seq datasets (*26*). Briefly, we predicted the gene expression of target cell type in dataset A with the gene expression of all cell types in dataset B, and vice versa, and then multiplied the resulting βs to determine the matching of cell types between the two data sets with high specificity. To calculate gene level accessibility scores from ATAC data, we summed the accessibility over gene bodies extended by 2 kb upstream of their TSS. In addition to determining the NNLS score for each cell type or cluster, accessibility close to known cell type marker genes [described in (*16*)] was inspected for each cluster in each tissue (summed over all cells in that cluster). Clusters that had a high score in the NNLS and/or clear specific accessibility at matching marker genes were annotated accordingly. Clusters without strong NNLS signal and weaker or less specific marker gene accessibility received a less confident annotation. Clusters with no NNLS signal and no or only uninformative marker gene expression were left unannotated. In some cases, several Louvain clusters received the same cell type annotation within a tissue and were merged accordingly for downstream analyses.

The same strategy was applied to transfer labels from adult or fetal expression data of human cortex. Processed gene-by-cell matrices were downloaded from the Allen Brain Map website for the adult data and from (*87*) for the fetal data.

Determining maternal contribution to cell types

To identify cell types with a lower fraction of Y chromosomal reads within a tissue, we selected all individual tissue samples with at least 800 cells and subsampled each cell type to 150 cells. For each tissue sample, we then calculated the ratio of Y chromosome over

autosome reads for these cells. As an additional line of evidence, we also used souporcell (*28*), a tool recently developed for clustering cells based on genotypes without a priori knowledge of individual genotypes.

Specificity scores

Cell type–specificity scores for each site–cell type pair were calculated by using Jensen-Shannon divergence as previously described (*13*). A list of the top 10,000 most specific peaks per cell type is provided in data file S4. Similarly, we calculated tissue specificity scores for cross-tissue analysis of shared cell types (fig. S9C).

Motif enrichments

We generated a binarized peak-by-motif matrix by identifying occurrences of motifs from the JASPAR vertebrate motif database (*106*) in each peak at a *P* value threshold of 10^{-7} using GC matched background nucleotide composition. A matrix of motif-by-cell counts was obtained by multiplying the peak-by-cell matrix with the peak-by-motif matrix. We downsampled the dataset so that a maximum of 800 cells per cell type, including unannotated clusters, were included to reduce computational cost and to reduce overrepresentation of very abundant cell types and tissues in computing enrichments in downstream steps. For each annotation, we then performed a negative binomial regression, predicting total motif counts using two input variables: an indicator column for the annotation as the main variable of interest and log(total number of nonzero entries in input peak matrix) for each cell as a covariate. We used the coefficient for the annotation indicator column and the intercept to estimate the fold change of the motif count of the annotation of interest relative to cells from all other annotations – exp(intercept + annotation_coefficient)/exp(intercept). This test was performed for all motifs in all groups, and *P* values were corrected by using the Benjamini Hochberg procedure.

For de novo motif finding, the top 2000 sites were selected by specificity score for each cell type (data file S4). We then ran homer (v4.11) using findMotifsGenome.pl with the specification -noknown -cpg (*48*). Matches to known motifs and scores were obtained from the standard homer output.

Trajectory analysis

Cells were subjected to trajectory analysis with Monocle3, similar to as previously described (*26*, *55*, *63*). When determining sites that changed over pseudotime, 10% of accessible sites (~50,000) were sampled to reduce computational complexity. Motifs enriched in sites with significant changes over pseudotime were determined by use of a logistic regression model predicting accessibility trends (opening or closing) with the presence or absence of JASPAR motifs called in the peak-by-motif matrix. For comparison with excitatory neuron dynamics in sci-RNA-seq data, excitatory neurons were extracted and compared with single-cell RNA-seq data from (*107*) with a previously described NNLS-based matching technique to identify progenitor and mature neuron populations (*26*). Pseudotime was determined with Monocle3, and expression of the gene of interest was normalized by size factor in each single cell, then log-transformed, scaled, and plotted over pseudotime.

Cicero models

Cicero coaccessibility analysis was performed for each of the 54 main cell types in each tissue (101 cell type–tissue combinations). Because coaccessibility scores are sensitive to false positives because of residual inter–cell type doublets driven by imperfect tissue dissociation, tissues were subjected to stringent doublet filtering on a per-tissue basis. To this end, the top 5000 specific peaks were selected per cell type (peak modules); each cell type was subclustered by using the union of these peak modules and a higher resolution, and subclusters with high accessibility in a peak module not matching the cell type in question were excluded. After applying this strategy, 91% of all cells were retained and subjected to coaccessibility score analysis. This stringently filtered dataset of 720,613 cells is provided at (*15*) and on the Gene Expression Omnibus (GEO) (GSE149683). The Cicero R package for Monocle3 (version 1.3.4.5) (*63*) was used to generate coaccessibility scores for each pair of sites within 500 kb in the linear genome and accessibility in at least 10 cells. Cicero gene activity scores were also generated by using default parameters.

Heritability enrichments

Heritability enrichments were calculated similar to our previous work (*13*). Our input set of peaks for each annotated cell type (without any distinction between the same annotation occurring across different tissues) were the top 10,000 peaks as ranked by specificity score within this set of annotations. UKBB traits were downloaded from (*108*). Only traits with an estimated heritability of 0.01 or higher were carried forward for analysis. *P* values were calculated from *z* scores assigned to coefficients (assuming two-sided test), and coefficients were divided by the average per-SNP heritability for the trait associated with a given test, producing scaled coefficients. Tests were corrected for multiple hypothesis testing by using the Benjamini-Hochberg method, and only tests with a *q* value of 0.2 or lower were deemed to be significant.

Testing for allelic imbalance

In order to identify genetic variation influencing chromatin accessibility levels, we used the "Quantitative Allele-Specific Analysis of Reads" (QuASAR) package in R (*68*). All reads from the selected samples were first remapped with HISAT2 (*109*), a SNP-aware aligner, to remove any positions showing an allelic mapping bias. Properly mapping reads were then sorted into individual bam files per cell type per individual. For each cell type, we generated a pileup file using Samtools (*110*) and considering all common SNPs from dbSNP144 (*111*). After filtering, we ran the analysis on all cell types from a sample at the same time using default parameters. The resulting *P* values were then FDR-corrected by using the Benjamini and Hochberg method (*112*) for each sample independently. Any *q* values less than 0.2 were considered significant. The method we used does not require prior genotyping, but because we had genotyped the individuals, we were able to confirm that sites inferred to be heterozygous in our analysis were also identified as heterozygous on the genotyping array: 2253 of 2254 (99.96%) overlapping SNPs for one individual and 6059 of 6065 (99.90%) for the other individual.

UMAP of sites

To reduce computational complexity, we used the peak-by-cell matrix based on 88,983 subsampled cells (subsampling up to 800 cells per cell type in each tissue, including unannotated clusters) and filtered out peaks smaller than 400 bp because these were less conserved and found in fewer cells. We transposed the resulting peak-by-cell matrix (447,879 by 88,983) and proceeded with Louvain clustering and UMAP visualization as above. To determine overlap of sites with regions bound by CTCF in GM12878, we downloaded the ChIP-seq peak locations from ENCODE (*56*, *57*). To determine overlap with looping anchors from Hi-C data in GM12878, we compared with loop annotations returned by the loop-calling algorithm HiCCUPS (*78*).

Comparison with mouse ATAC atlas

To compare peak coordinates, mouse peaks from (*13*) were lifted over from mm9 to hg19 by using the University of California, Santa Cruz liftOver tool (*113*). To compare cell type–specific motif enrichments, a peak-by-motif matrix was generated for all mouse peaks and cell types from (*13*) by using the same motif database and *P* value cutoff, and motif enrichment analysis was conducted as described above.

Comparison with adult single-cell ATAC-seq data

Comparison with adult single-cell ATAC-seq data was performed on peak-by-cell matrices. For blood (63,883 nuclei from adult bone marrow and peripheral blood profiled on the 10X platform), this matrix was directly downloaded (*84*); for cortex (12,557 nuclei from post-mortem human brain profiled with sn-ATAC-seq), it was generated from the snap file and peak calls

provided by the authors (*85*, *104*). Fetal and adult peaks from both master lists were intersected, and 100,000 overlapping peaks were randomly sampled from the adult peak set to reduce computational complexity. Then the fetal data, downsampled to 1500 cells per Louvain cluster, was rescored on the basis of these peaks. Subsequently, the datasets were integrated by using integration anchors and the Signac package (*23*, *114*). For motif analysis, a peak-by-motif matrix was generated for all peaks in the adult blood dataset by using the same motif database and *P* value cutoff as described above, and motif enrichment calculation was carried out accordingly.

REFERENCES AND NOTES

1. D. A. Cusanovich *et al*., Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015). doi: 10.1126/science.aab1601; pmid: 25953818
2. J. Cao *et al*., Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017). doi: 10.1126/science.aam8940; pmid: 28818938
3. V. Ramani *et al*., Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017). doi: 10.1038/nmeth.4155; pmid: 28135255
4. R. M. Mulqueen *et al*., Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018). doi: 10.1038/nbt.4112; pmid: 29644997
5. Y. Yin *et al*., High-Throughput Single-Cell Sequencing with Linear Amplification. *Mol. Cell* **76**, 676–690.e10 (2019). doi: 10.1016/j.molcel.2019.08.002; pmid: 31495564
6. J. Cao *et al*., Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018). doi: 10.1126/science.aau0730; pmid: 30166440
7. S. A. Vitak *et al*., Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017). doi: 10.1038/nmeth.4154; pmid: 28135258
8. A. B. Rosenberg *et al*., Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018). doi: 10.1126/science.aam8999; pmid: 29545511
9. C. A. Lareau *et al*., Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019). doi: 10.1038/s41587-019-0147-6; pmid: 31235917
10. Q. Wang *et al*., CoBATCH for High-Throughput Single-Cell Epigenomic Profiling. *Mol. Cell* **76**, 206–216.e7 (2019). doi: 10.1016/j.molcel.2019.07.015; pmid: 31471188
11. J. Cao, W. Zhou, F. Steemers, C. Trapnell, J. Shendure, Sci-fate characterizes the dynamics of gene expression in single cells. *Nat. Biotechnol.* **38**, 980–988 (2020). doi: 10.1038/s41587-020-0480-9; pmid: 32284584
12. C. Zhu *et al*., An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019). doi: 10.1038/s41594-019-0323-x; pmid: 31695190
13. D. A. Cusanovich *et al*., A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e18 (2018). doi: 10.1016/j.cell.2018.06.052; pmid: 30078704
14. S. Behjati, S. Lindsay, S. A. Teichmann, M. Haniffa, Mapping human development at single-cell resolution. *Development* **145**, dev152561 (2018). doi: 10.1242/dev.152561; pmid: 29439135
15. descartes.brotmanbaty.org
16. J. Cao *et al*., A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020). doi: 10.1126/science.aba7721
17. F. Yue *et al*., A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014). doi: 10.1038/nature13992; pmid: 25409824
18. J. D. Bloom, Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ* **6**, e5578 (2018). doi: 10.7717/peerj.5578; pmid: 30202659
19. H. Chen *et al*., Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019). doi: 10.1186/s13059-019-1854-5; pmid: 31739806
20. A. J. Hill, Andrew John Hill Blog; http://andrewjohnhill.com/blog/2019/05/06/dimensionality-reduction-for-scatac-data.
21. I. Korsunsky *et al*., Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019). doi: 10.1038/s41592-019-0619-0; pmid: 31740819
22. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018). doi: 10.21105/joss.00861
23. T. Stuart *et al*., Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019). doi: 10.1016/j.cell.2019.05.031; pmid: 31178118
24. B. B. Lake *et al*., Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018). doi: 10.1038/nbt.4038; pmid: 29227469
25. L. T. Graybuck *et al*., Prospective, brain-wide labeling of neuronal subclasses with enhancer-driven AAVs. *bioRxiv* (2019). doi: 10.1101/525014
26. J. Cao *et al*., The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019). doi: 10.1038/s41586-019-0969-x; pmid: 30787437
27. H. Suryawanshi *et al*., A single-cell survey of the human first-trimester placenta and decidua. *Sci. Adv.* **4**, eaau4788 (2018). doi: 10.1126/sciadv.aau4788; pmid: 30402542
28. H. Heaton *et al*., Souporcell: Robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020). doi: 10.1038/s41592-020-0820-1; pmid: 32366989
29. H. Iwasaki, K. Akashi, Myeloid lineage commitment from the hematopoietic stem cell. *Immunity* **26**, 726–740 (2007). doi: 10.1016/j.immuni.2007.06.004; pmid: 17582345
30. A. Arthur *et al*., Twist-1 Enhances Bone Marrow Mesenchymal Stromal Cell Support of Hematopoiesis by Modulating CXCL12 Expression. *Stem Cells* **34**, 504–509 (2016). doi: 10.1002/stem.2265; pmid: 26718114
31. S. J. Renaud, K. Kubota, M. A. K. Rumi, M. J. Soares, The FOS transcription factor family differentially controls trophoblast migration and invasion. *J. Biol. Chem.* **289**, 5025–5039 (2014). doi: 10.1074/jbc.M113.523746; pmid: 24379408
32. A.-M. Bamberger *et al*., Expression pattern of the activating protein-1 family of transcription factors in the human placenta. *Mol. Hum. Reprod.* **10**, 223–228 (2004). doi: 10.1093/molehr/gah011; pmid: 14985474
33. J. Palis, Primitive and definitive erythropoiesis in mammals. *Front. Physiol.* **5**, 3 (2014). doi: 10.3389/fphys.2014.00003; pmid: 24478716
34. A. K. Hennig, G.-H. Peng, S. Chen, Regulation of photoreceptor gene expression by Crx-associated transcription factor network. *Brain Res.* **1192**, 114–133 (2008). doi: 10.1016/j.brainres.2007.06.036; pmid: 17662965
35. B. L. Black, E. N. Olson, Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. *Annu. Rev. Cell Dev. Biol.* **14**, 167–196 (1998). doi: 10.1146/annurev.cellbio.14.1.167; pmid: 9891782
36. Z. Niu *et al*., Serum response factor orchestrates nascent sarcomerogenesis and silences the biomineralization gene program in the heart. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 17824–17829 (2008). doi: 10.1073/pnas.0805491105; pmid: 19004760
37. R. Kageyama, H. Shimojo, T. Ohtsuka, Dynamic control of neural stem cells by bHLH factors. *Neurosci. Res.* **138**, 12–18 (2019). doi: 10.1016/j.neures.2018.09.005; pmid: 30227160
38. G. Wilkinson, D. Dennis, C. Schuurmans, Proneural genes in neocortical development. *Neuroscience* **253**, 256–273 (2013). doi: 10.1016/j.neuroscience.2013.08.029; pmid: 23999125
39. M. Zou, S. Li, W. H. Klein, M. Xiang, Brn3a/Pou4f1 regulates dorsal root ganglion sensory neuron specification and axonal projection into the spinal cord. *Dev. Biol.* **364**, 114–127 (2012). doi: 10.1016/j.ydbio.2012.01.021; pmid: 22326227
40. M. G. De Vas *et al*., Hnf1b controls pancreas morphogenesis and the generation of Ngn3^{+} endocrine progenitors. *Development* **142**, 871–882 (2015). doi: 10.1242/dev.110759; pmid: 25715395
41. A. Desgrange *et al*., HNF1B controls epithelial organization and cell polarity during ureteric bud branching and collecting duct morphogenesis. *Development* **144**, 4704–4719 (2017). doi: 10.1242/dev.154336; pmid: 29158444
42. K. Vázquez-Arreguín, D. Tantin, The Oct1 transcription factor and epithelial malignancies: Old protein learns new tricks. *Biochim. Biophys. Acta* **1859**, 792–804 (2016). doi: 10.1016/j.bbagrm.2016.02.007; pmid: 26877236
43. D. McClellan *et al*., Growth factor independence 1B-mediated transcriptional repression and lineage allocation require lysine-specific demethylase 1-dependent recruitment of the BHC complex. *Mol. Cell. Biol.* **39**, e00020–e19 (2019). doi: 10.1128/MCB.00020-19; pmid: 30988160
44. M. Oukka *et al*., The transcription factor NFAT4 is involved in the generation and survival of T cells. *Immunity* **9**, 295–304 (1998). doi: 10.1016/S1074-7613(00)80612-3; pmid: 9768749
45. Z. Li *et al*., Serine 574 phosphorylation alters transcriptional programming of FOXO3 by selectively enhancing apoptotic gene expression. *Cell Death Differ.* **23**, 583–595 (2016). doi: 10.1038/cdd.2015.125; pmid: 26470730
46. A. Jankowski, J. Tiuryn, S. Prabhakar, Romulus: Robust multi-state identification of transcription factor binding sites from DNase-seq data. *Bioinformatics* **32**, 2419–2426 (2016). doi: 10.1093/bioinformatics/btw209; pmid: 27153645
47. Z. F. Chen, A. J. Paquette, D. J. Anderson, NRSF/REST is required in vivo for repression of multiple neuronal target genes during embryogenesis. *Nat. Genet.* **20**, 136–142 (1998). doi: 10.1038/2431; pmid: 9771705
48. S. Heinz *et al*., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010). doi: 10.1016/j.molcel.2010.05.004; pmid: 20513432
49. Y. Lavin *et al*., Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment. *Cell* **159**, 1312–1326 (2014). doi: 10.1016/j.cell.2014.11.018; pmid: 25480296
50. D. B. Thomas, J. M. Yoffey, Human foetal haematopoiesis. II. Hepatic haematopoiesis in the human foetus. *Br. J. Haematol.* **10**, 193–197 (1964). doi: 10.1111/j.1365-2141.1964.tb00694.x; pmid: 14141618
51. L. O. Jacobson, E. L. Simmons, E. K. Marks, J. H. Eldredge, Recovery from radiation injury. *Science* **113**, 510–511 (1951). doi: 10.1126/science.113.2940.510; pmid: 14828383
52. B. K. Tusi *et al*., Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018). doi: 10.1038/nature25741; pmid: 29466336
53. V. P. Schulz *et al*., A unique epigenomic landscape defines human erythropoiesis. *Cell Rep.* **28**, 2996–3009.e7 (2019). doi: 10.1016/j.celrep.2019.08.020; pmid: 31509757
54. T. J. Satchwell, Erythrocyte invasion receptors for *Plasmodium falciparum*: New and old. *Transfus. Med.* **26**, 77–88 (2016). doi: 10.1111/tme.12280; pmid: 26862042
55. C. Trapnell *et al*., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014). doi: 10.1038/nbt.2859; pmid: 24658644
56. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). doi: 10.1038/nature11247; pmid: 22955616
57. C. A. Davis *et al*., The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46** (D1), D794–D801 (2018). doi: 10.1093/nar/gkx1081; pmid: 29126249
58. Roadmap Epigenomics Consortium, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
59. S. De Val, B. L. Black, Transcriptional control of endothelial cell development. *Dev. Cell* **16**, 180–195 (2009). doi: 10.1016/j.devcel.2009.01.014; pmid: 19217421
60. A. V. Shah, G. M. Birdsey, A. M. Randi, Regulation of endothelial homeostasis, vascular development and angiogenesis by the transcription factor ERG. *Vascul. Pharmacol.* **86**, 3–13 (2016). doi: 10.1016/j.vph.2016.05.003; pmid: 27208692
61. D. Siripin *et al*., Transdifferentiation of erythroblasts to megakaryocytes using FLI1 and ERG transcription factors. *Thromb. Haemost.* **114**, 593–602 (2015). doi: 10.1160/TH14-12-1090; pmid: 26063314
62. O. Goldman *et al*., Endoderm generates endothelial cells during liver development. *Stem Cell Reports* **3**, 556–565 (2014). doi: 10.1016/j.stemcr.2014.08.009; pmid: 25358784
63. H. A. Pliner *et al*., Cicero Predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018). doi: 10.1016/j.molcel.2018.06.044; pmid: 30078726
64. D. Noordermeer, W. de Laat, Joining the loops: Beta-globin gene regulation. *IUBMB Life* **60**, 824–833 (2008). doi: 10.1002/iub.129; pmid: 18767169
65. J. K. Pickrell, Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014). doi: 10.1016/j.ajhg.2014.03.004; pmid: 24702953

66. M. T. Maurano *et al*., Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012). doi: 10.1126/science.1222794; pmid: 22955828
67. H. K. Finucane *et al*., Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015). doi: 10.1038/ng.3404; pmid: 26414678
68. C. T. Harvey *et al*., QuASAR: Quantitative allele-specific analysis of reads. *Bioinformatics* **31**, 1235–1242 (2015). doi: 10.1093/bioinformatics/btu802; pmid: 25480375
69. A. Buniello *et al*., The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47** (D1), D1005–D1012 (2019). doi: 10.1093/nar/gky1120; pmid: 30445434
70. G. Kichaev *et al*., Leveraging polygenic functional enrichment to improve GWAS power. *Am. J. Hum. Genet.* **104**, 65–75 (2019). doi: 10.1016/j.ajhg.2018.11.008; pmid: 30595370
71. N. Tanimura *et al*., GATA/heme multi-omics reveals a trace metal-dependent cellular differentiation mechanism. *Dev. Cell* **46**, 581–594.e4 (2018). doi: 10.1016/j.devcel.2018.07.022; pmid: 30122630
72. B. Zhao *et al*., Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat. Genet.* **51**, 1637–1644 (2019). doi: 10.1038/s41588-019-0516-6; pmid: 31676860
73. F. Tissir, A. M. Goffinet, Reelin and brain development. *Nat. Rev. Neurosci.* **4**, 496–505 (2003). doi: 10.1038/nrn1113; pmid: 12778121
74. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, chromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017). doi: 10.1038/nmeth.4401; pmid: 28825706
75. M. W. Dorrity, L. M. Saunders, C. Queitsch, S. Fields, C. Trapnell, Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nat. Commun.* **11**, 1537 (2020). doi: 10.1038/s41467-020-15351-4; pmid: 32210240
76. E. Eisenberg, E. Y. Levanon, Human housekeeping genes, revisited. *Trends Genet.* **29**, 569–574 (2013). doi: 10.1016/j.tig.2013.05.010; pmid: 23810203
77. J. R. Dixon *et al*., Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012). doi: 10.1038/nature11082; pmid: 22495300
78. S. S. P. Rao *et al*., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014). doi: 10.1016/j.cell.2014.11.021; pmid: 25497547
79. T. H. Kim *et al*., Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007). doi: 10.1016/j.cell.2006.12.048; pmid: 17382889
80. R. M. Gittelman *et al*., Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Res.* **25**, 1245–1255 (2015). doi: 10.1101/gr.192591.115; pmid: 26104583
81. A. Visel, S. Minovitsky, I. Dubchak, L. A. Pennacchio, VISTA Enhancer Browser—A database of tissue-specific human enhancers. *Nucleic Acids Res.* **35** (Database), D88–D92 (2007). doi: 10.1093/nar/gkl822; pmid: 17130149
82. D. C. King *et al*., Finding cis-regulatory elements using comparative genomics: Some lessons from ENCODE data. *Genome Res.* **17**, 775–786 (2007). doi: 10.1101/gr.5592107; pmid: 17567996
83. E. Montecino-Rodriguez, K. Dorshkind, Formation of B-1 B cells from neonatal B-1 transitional cells exhibits NF-κB redundancy. *J. Immunol.* **187**, 5712–5719 (2011). doi: 10.4049/jimmunol.1102416; pmid: 22031760
84. A. T. Satpathy *et al*., Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019). doi: 10.1038/s41587-019-0206-z; pmid: 31375813
85. C. Luo *et al*., Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv* 873398 [Preprint] 12 December 2019; .doi: 10.1101/2019.12.11.873398
86. Cell Types Database, RNA-Seq Data—brain-map.org; https://portal.brain-map.org/atlases-and-data/rnaseq#Human_Cortex.
87. D. Polioudakis *et al*., A single-cell transcriptomic atlas of human neocortical development during mid-gestation. *Neuron* **103**, 785–801.e8 (2019). doi: 10.1016/j.neuron.2019.06.011; pmid: 31303374
88. A. Bery, B. Martynoga, F. Guillemot, J.-S. Joly, S. Rétaux, Characterization of enhancers active in the mouse embryonic cerebral cortex suggests Sox/Pou cis-regulatory logics and heterogeneity of cortical progenitors. *Cereb. Cortex* **24**, 2822–2834 (2014). doi: 10.1093/cercor/bht126; pmid: 23720416
89. S. E. McCandless, J. W. Brunger, S. B. Cassidy, The burden of genetic disease on inpatient care in a children's hospital. *Am. J. Hum. Genet.* **74**, 121–127 (2004). doi: 10.1086/381053; pmid: 14681831
90. M. H. Wojcik *et al*., Genetic disorders and mortality in infancy and early childhood: Delayed diagnoses and missed opportunities. *Genet. Med.* **20**, 1396–1404 (2018). doi: 10.1038/gim.2018.17; pmid: 29790870
91. W. Meuleman *et al*., Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020). doi: 10.1038/s41586-020-2559-3; pmid: 32728217
92. M. R. Corces *et al*., Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016). doi: 10.1038/ng.3646; pmid: 27526324
93. G. J. Pruijn, W. van Driel, P. C. van der Vliet, Nuclear factor III, a novel sequence-specific DNA-binding protein from HeLa cells stimulating adenovirus DNA replication. *Nature* **322**, 656–659 (1986). doi: 10.1038/322656a0; pmid: 3748145
94. D. Tantin, C. Schild-Poulter, V. Wang, R. J. G. Haché, P. A. Sharp, The octamer binding transcription factor Oct-1 is a stress sensor. *Cancer Res.* **65**, 10750–10758 (2005). doi: 10.1158/0008-5472.CAN-05-2399; pmid: 16322220
95. V. E. H. Wang, T. Schmidt, J. Chen, P. A. Sharp, D. Tantin, Embryonic lethality, decreased erythropoiesis, and defective octamer-dependent promoter activation in Oct-1-deficient mice. *Mol. Cell. Biol.* **24**, 1022–1032 (2004). doi: 10.1128/MCB.24.3.1022-1032.2004; pmid: 14729950
96. D. Lee *et al*., A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015). doi: 10.1038/ng.3331; pmid: 26075791
97. S. Domcke, A. J. Hill, R. M. Daza, C. Trapnell, D. A. Cusanovich, J. Shendure, sci-ATAC-seq3. protocols.io (2020). doi: 10.17504/protocols.io.be8mjhu6
98. M. R. Corces *et al*., An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017). doi: 10.1038/nmeth.4396; pmid: 28846090
99. A. Saunders, L. J. Core, C. Sutcliffe, J. T. Lis, H. L. Ashe, Extensive polymerase pausing during *Drosophila* axis patterning enables high-level and pliable transcription. *Genes Dev.* **27**, 1146–1158 (2013). doi: 10.1101/gad.215459.113; pmid: 23699410
100. S. Domcke, A. Hill, shendurelab/human-atac: First release of sci-ATAC-seq3 demultiplexing. Zenodo (2020). doi: 10.5281/zenodo.4012524
101. Y. Zhang *et al*., Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008). doi: 10.1186/gb-2008-9-9-r137; pmid: 18798982
102. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst.* **8**, 281–291.e9 (2019). doi: 10.1016/j.cels.2018.11.005; pmid: 30954476
103. C. Bravo González-Blas *et al*., cisTopic: Cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019). doi: 10.1038/s41592-019-0367-1; pmid: 30962623
104. R. Fang *et al*., Fast and accurate clustering of single cell epigenomes reveals cis-regulatory elements in rare cell types. *bioRxiv* 615179 [Preprint] 13 May 2019).
105. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018). doi: 10.1038/nbt.4096; pmid: 29608179
106. O. Fornes *et al*., JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020). doi: 10.1093/nar/gkz1001; pmid: 31701148
107. A. Zeisel *et al*., Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014.e22 (2018). doi: 10.1016/j.cell.2018.06.021; pmid: 30096314
108. https://data.broadinstitute.org/alkesgroup/UKBB/UKBB_409K
109. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019). doi: 10.1038/s41587-019-0201-4; pmid: 31375807
110. H. Li *et al*., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). doi: 10.1093/bioinformatics/btp352; pmid: 19505943
111. S. T. Sherry *et al*., dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001). doi: 10.1093/nar/29.1.308; pmid: 11125122
112. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995). doi: 10.1111/j.2517-6161.1995.tb02031.x
113. W. J. Kent *et al*., The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002). doi: 10.1101/gr.229102; pmid: 12045153
114. T. Stuart, timoast/signac. GitHub (2020); https://github.com/timoast/signac.
115. Broad Institute, *Picard Tools*; http://broadinstitute.github.io/picard.
116. J. A. Castro-Mondragon, S. Jaeger, D. Thieffry, M. Thomas-Chollier, J. van Helden, RSAT matrix-clustering: Dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.* **45**, e119 (2017). doi: 10.1093/nar/gkx314; pmid: 28591841
117. L. Navarro-Núñez, S. A. Langan, G. B. Nash, S. P. Watson, The physiological and pathophysiological roles of platelet CLEC-2. *Thromb. Haemost.* **109**, 991–998 (2013). doi: 10.1160/TH13-01-0060; pmid: 23572154

ACKNOWLEDGMENTS

We thank past and present members of the Shendure, Trapnell, and Cusanovich laboratories; A. Adey; the Brotman Baty Institute for Precision Medicine (BBI) Advanced Technology Laboratory; and the University of Washington Birth Defects Research Laboratory. We thank C. Spurrell, D. Ahrendsen, and R. Blecher for reviewing the sci-ATAC-seq3 protocol. We thank C. Cenik and Z. Ma from M. Snyder's laboratory for providing the CH12.LX cell line. We thank R. Pique-Regi for advice on implementing QuASAR. We acknowledge the ENCODE Consortium for access to their data. **Funding:** S.D. was supported by an EMBO Long-Term Fellowship and a HFSP Long-Term Fellowship. D.A.C. was supported in part by an NHLBI fellowship (T32HL007828). Aspects of this work were supported by funding from the BBI, the Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to J.S. and C.T.), The Chan Zuckerberg Initiative (to C.T.) and the NIH (HD000836 to I.A.G.). J.S. is an investigator of the Howard Hughes Medical Institute. **Authors contributions:** R.M.D., S.D., and D.A.C. developed techniques and performed sci-ATAC-seq3 experiments with assistance from F.Z., D.P., F.J.S., and J.H.M.; D.R.O. performed tissue collection and nuclei extraction under supervision of D.D. and I.A.G.; S.D. and A.J.H. analyzed the data with assistance from D.A.C. and H.A.P.; J.C. provided sci-RNA-seq data; K.A.A. performed genotyping; M.A.Z. developed the website, with assistance from S.D.; S.D., D.A.C., and J.S. wrote the manuscript with input from all co-authors; J.S., D.A.C., and C.T. supervised the project. **Competing interests:** D.P., F.Z., and F.J.S. declare competing financial interests in the form of stock ownership and paid employment by Illumina. J.S. has competing financial interests (paid consulting and/or equity) with Guardant Health, Maze Therapeutics, Camp4 Therapeutics, Nanostring, Phase Genomics, Adaptive Biotechnologies, and Stratos Genomics. One or more embodiments of one or more patents and patent applications filed by Illumina and the University of Washington may encompass the methods, reagents, and data disclosed in this manuscript. **Data and materials availability:** In contrast with previous versions of sci-ATAC-seq, custom Tn5 is not required for sci-ATAC-seq3. A detailed version of the sci-ATAC-seq3 protocol is available on protocols.io (*97*). A demultiplexing script and tutorial are provided on Zenodo at (100). Supplementary data files S1 to S7 are provided at GEO (accession no. GSE149683). Raw data are provided at dbGaP (accession no. phs002003.v1.p1). Processed data are available at GEO (accession no. GSE149683 (both unfiltered and peak module-based filtered dataset)). An interactive version of the peak module–based filtered dataset facilitates the exploration of these data by tissue, cell type, locus, or motif at descartes.brotmanbaty.org (*15*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/eaba7612/suppl/DC1
Materials and Methods
Figs. S1 to S19
References (*118*–*124*)
Tables S1 to S7

View/request a protocol for this paper from *Bio-protocol*.

3 January 2020; accepted 10 September 2020
10.1126/science.aba7612

RESEARCH ARTICLE SUMMARY

PLANT SCIENCE

Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions

Jos R. Wendrich*, BaoJun Yang*, Niels Vandamme*, Kevin Verstaen*, Wouter Smet, Celien Van de Velde, Max Minne, Brecht Wybouw, Eliana Mor, Helena E. Arents, Jonah Nolf, Julie Van Duyse, Gert Van Isterdael, Steven Maere, Yvan Saeys†, Bert De Rybel†

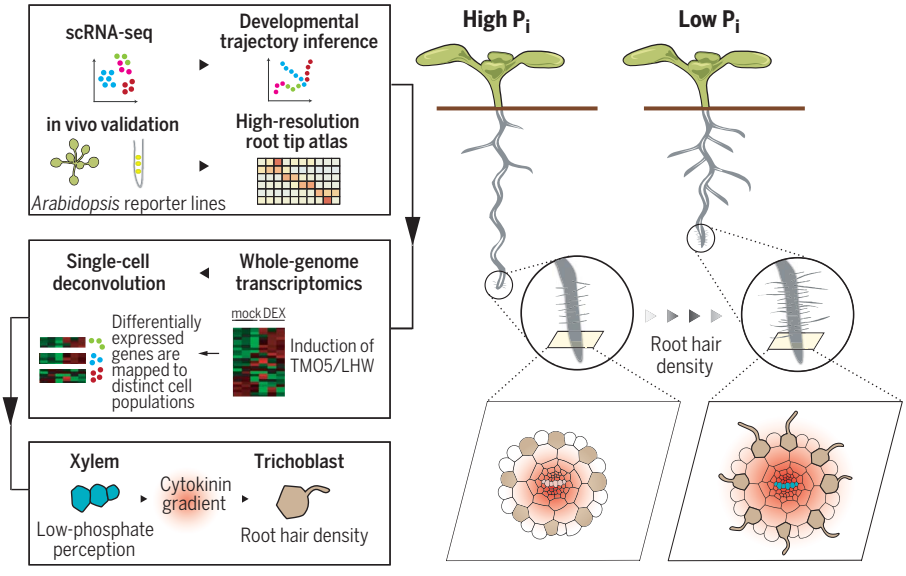
INTRODUCTION: In plants, vascular tissues serve a dual function by providing both structural support and transport of water, nutrients, hormones, and other signaling molecules throughout the plant body. Proliferation of vascular tissues is, in part, controlled by the TARGET OF MONOPTEROS 5/LONESOME HIGHWAY (TMO5/LHW) transcription factor complex and involves intricate cell communication through mobile factors. The activity of this heterodimer complex is limited to young xylem cells and is required and sufficient to control vascular cell proliferation by inducing expression of rate-limiting enzymes in the cytokinin biosynthetic pathway. Because cytokinin is mobile, the molecular and developmental responses to this hormone signaling cascade are likely to have an effect in various cell types surrounding the xylem and perhaps even outside the vascular bundle. To what extent this is the case and how this controls developmental processes other than vascular proliferation remain unknown.

RATIONALE: By intersecting a high-resolution single-cell gene expression atlas with TMO5/LHW target genes, we probed the tissue-specific distribution of transcriptional responses upon induction of this transcription factor complex in the *Arabidopsis* root meristem. We next used a combination of genetic and molecular tools to validate the existence and importance of a cytokinin-dependent signaling mechanism bridging the inner vascular tissues and the outer epidermis cell layers.

RESULTS: We generated a single-cell gene expression atlas that represents all known cell types and sub–cell types of the *Arabidopsis* root apical meristem. Cell-type annotations and developmental trajectories were validated through expression of known marker genes, newly generated promoter reporter lines, and analysis of the endoreduplication status in each cell cluster. We next intersected this expression atlas with a set of known target genes of the vascular TMO5/LHW transcription factor complex and found that a large portion of the target genes are expressed in, and are even restricted to, the outermost root hair cells. This suggests a possible involvement of the TMO5/LHW complex in root hair development. In line with this hypothesis, misexpression of TMO5 and LHW in the entire root meristem was found to increase root hair density, a phenotype resembling wild-type roots grown under phosphate-limiting conditions. Moreover, the response of epidermal cells to phosphate deficit was shown to be TMO5-dependent. Phosphate-limiting conditions were found to induce *TMO5* expression, and genetically increasing *TMO5* expression only in the vascular bundle is sufficient to trigger epidermal responses. These results indicate that the TMO5-dependent increase in root hair density under low-phosphate conditions occurs through a cell nonautonomous effect.

TMO5/LHW-dependent vascular cytokinin biosynthesis was found to be sufficient to increase epidermal root hair density because mutants with reduced cytokinin levels are less sensitive to the effects of phosphate-limiting conditions on root hair density. Furthermore, phosphate-limiting conditions increase cytokinin signaling in epidermal cells of wild-type roots, and treatment with exogenous cytokinin can restore root hair densities back to wild-type levels in the absence of TMO5 activity. This suggests that cytokinin is the mobile signal linking TMO5/LHW activity in vascular cells to root hair density changes in response to phosphate deficit. The TMO5 and cytokinin-dependent root hair density increase under low-phosphate conditions was found to originate not only from changes in epidermal cell length but also from changes in cell identity because both phosphate-limiting conditions and exogenous cytokinin treatment were shown to scramble identities of the otherwise highly organized epidermal cells.

CONCLUSION: Through its effect on cytokinin biosynthesis, the vascular TMO5/LHW heterodimer complex controls epidermal root hair density by modifying cell length and cell identity. We hypothesize that phosphate-limiting conditions may trigger increased auxin signaling in xylem cells, inducing activity of the TMO5/LHW pathway and downstream local cytokinin biosynthesis. Cytokinin may then diffuse outward from the vasculature to direct both length and fate of the outer trichoblast cells. This hormone signaling cascade spans multiple tissue layers, allowing roots to efficiently forage the soil for phosphate. ■



Vascular transcription factors guide epidermal responses. A validated high-resolution single-cell gene expression atlas of the *Arabidopsis* root was intersected with TMO5/LHW target genes, uncovering an enrichment of epidermal-restricted expression patterns. By activating local cytokinin biosynthesis, the vascular TMO5/LHW complex was shown to regulate epidermal root hair density in response to the availability of phosphate. scRNA-seq, single-cell RNA-sequencing; mock, untreated condition; DEX, dexamethasone; P_i, inorganic phosphate.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. Email: beryb@psb.vib-ugent.be (B.D.R.); yvan.saeys@irc.vib-ugent.be (Y.S.)
Cite this article as J. R. Wendrich *et al.*, *Science* **370**, eaay4970 (2020). DOI: 10.1126/science.aay4970

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.aay4970

RESEARCH ARTICLE

PLANT SCIENCE

Vascular transcription factors guide plant epidermal responses to limiting phosphate conditions

Jos R. Wendrich[1,2]*, BaoJun Yang[1,2]*, Niels Vandamme[3,4]*, Kevin Verstaen[3,4]*, Wouter Smet[1,2], Celien Van de Velde[1,2], Max Minne[1,2], Brecht Wybouw[1,2], Eliana Mor[1,2], Helena E. Arents[1,2], Jonah Nolf[1,2], Julie Van Duyse[5,6], Gert Van Isterdael[5,6], Steven Maere[1,2], Yvan Saeys[3,4]†, Bert De Rybel[1,2]†

Optimal plant growth is hampered by deficiency of the essential macronutrient phosphate in most soils. Plant roots can, however, increase their root hair density to efficiently forage the soil for this immobile nutrient. By generating and exploiting a high-resolution single-cell gene expression atlas of *Arabidopsis* roots, we show an enrichment of TARGET OF MONOPTEROS 5/LONESOME HIGHWAY (TMO5/LHW) target gene responses in root hair cells. The TMO5/LHW heterodimer triggers biosynthesis of mobile cytokinin in vascular cells and increases root hair density during low-phosphate conditions by modifying both the length and cell fate of epidermal cells. Moreover, root hair responses in phosphate-deprived conditions are TMO5- and cytokinin-dependent. Cytokinin signaling links root hair responses in the epidermis to perception of phosphate depletion in vascular cells.

Vascular cell proliferation in plant roots is, in part, controlled by the heterodimer complex formed by TARGET OF MONOPTEROS 5 and LONESOME HIGHWAY (TMO5/LHW) (*1*–*7*). This complex is required and sufficient to control cell proliferation by inducing expression of the direct downstream target *LONELY GUY4* (*LOG4*) and its close homolog *LOG3* (*1*, *6*), which encode rate-limiting enzymes in the final step of the conversion of the phytohormone cytokinin into its bioactive form (*8*, *9*). The TMO5/LHW complex is limited to xylem cells, which produce cytokinin but are themselves insensitive to cytokinin. The xylem-produced cytokinin diffuses to neighboring procambium cells, where it promotes cell proliferation by induction of DOF-type transcription factors (*10*, *11*). Because the TMO5/LHW pathway induces production of cytokinin as a mobile intermediate that functions in neighboring cells, the target genes (*11*) in this hormone signaling cascade are likely to be expressed in various cell types surrounding the xylem and perhaps even outside of the vascular bundle. Here, we used single-cell RNA-sequencing (scRNA-seq) to probe the tissue-specific TMO5/LHW signaling output in *Arabidopsis* root meristems and found that this vascular heterodimer complex is required for the root hair responses to phosphate-deficit conditions. We show how cytokinin signaling links vascular perception of limiting phosphate to epidermal responses, which allows plants to efficiently forage the soil for this immobile macronutrient.

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium. [2]VIB Center for Plant Systems Biology, Ghent, Belgium. [3]Data Mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium. [4]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium. [5]VIB Flow Core, VIB Center for Inflammation Research, Ghent, Belgium. [6]Department of Biomedical Molecular Biology, Ghent University, Ghent, Belgium.
*These authors contributed equally to this work.
†Corresponding author. Email: beryb@psb.vib-ugent.be (B.D.R.); yvan.saeys@irc.vib-ugent.be (Y.S.)

Single-cell RNA-sequencing analysis

We generated a high-resolution scRNA-seq atlas of the wild-type *Arabidopsis* root tip (*12*–*16*), making use of the 10x Genomics Chromium technology (fig. S1A). After protoplast isolation, sorted cells were collected and processed for single-cell transcriptomics (see supplementary materials for details). A total population of 15,918 cells was recovered across three replicates and then filtered to retain 5145 high-quality cells with unique molecular identifier (UMI) counts >17,290 (fig. S1B). Taking into account only these high-quality cells, a total of 21,492 genes were detected in our root meristem dataset, covering nearly 80% of the genome, with a median expression of 6781 genes per cell and a mean of 208,937 reads per cell (Fig. 1A). Unsupervised clustering and *t*-distributed stochastic neighbor embedding (tSNE) projections were performed on the 5145 high-quality single cells, recovering distinct clusters of cells (Fig. 1A).

After quality control (see supplementary materials and fig. S1, C to E), cell-type annotation and cluster identification were performed by mapping the top 20 differentially expressed genes (DEGs) for each cluster (compared with the rest of the dataset) on a publicly available bulk RNA-seq dataset (*17*) of the *Arabidopsis* root. This resulted in an annotated dataset representing all major cell types in the root, including quiescent center cells (Fig. 1A and fig. S1F). The annotations were confirmed for all cell identities by the observed expression of key signature marker genes and in vivo expression of 41 newly generated promoter reporter lines (Fig. 1B, figs. S2 to S12, and table S1). Moreover, cells undergoing division were found in two specific subclusters (Fig. 1), indicating that their transcriptomes are more similar to each other than to the actual cell identity determinants of the transcriptome. The scRNA-seq dataset contains clusters of all cell types of the root meristem, which were identified according to predicted in vivo expression patterns and validated using a set of newly generated reporter lines.

Trajectory analysis establishes a blueprint of cell lineages

We next used trajectory analyses to refine identification of cell types and developmental transitions within each cell identity cluster. We first analyzed xylem (251 cells, 5%; fig. S3A) and phloem (388 cells, 8%; fig. S4A) cell lineages, because these undergo identity changes throughout development. Xylem initial cell lineages branch into proto- and metaxylem identities, which differ in their subsequent differentiation processes, including secondary cell wall generation (*18*). Phloem cell initials undergo several oriented divisions, generating lineages that branch to generate phloem procambium, sieve elements, and companion cells (*19*, *20*). The complexity of both these cell types was captured in the inferred trajectories, and gene expression patterns of the reporter lines validated inferred developmental trajectories and subcluster identities (figs. S3 and S4; see supplementary materials for details). Similar analyses validated trajectories for the procambium, pericycle, endodermis, cortex, epidermis, lateral root cap, and columella clusters (Fig. 1A and figs. S5 to S11).

Because many root cell types in *Arabidopsis* increase in ploidy with development (*21*), the developmental trajectories for these cell types should correspond to trajectories of increasing cellular endoreplication levels. Thus, to further validate our trajectories, we predicted the endoreplication state of each cell on the basis of the expression of a validated set of endoreplication markers (*21*) (see supplementary materials for details) and superimposed the predicted cell ploidy on the tSNE plot (Fig. 1C). Developmental trajectories of cortex, endodermis, pericycle, epidermis and atrichoblast, lateral root cap, and xylem clusters exhibit clear ploidy transitions (Fig. 1C), validating these trajectories and their orientation. For the procambium cell cluster, correspondence between developmental trajectories and ploidy was less evident. For the phloem cell lineage,

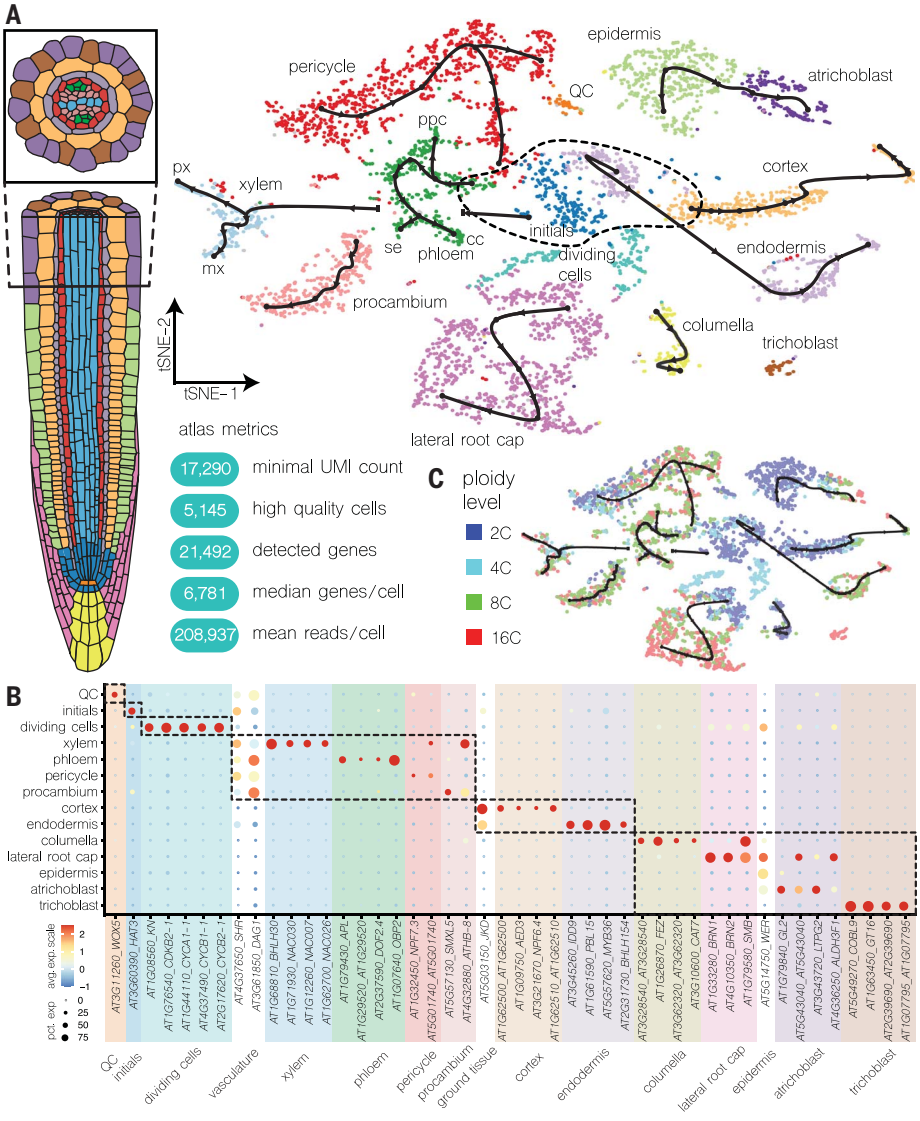


Fig. 1. Identification of *Arabidopsis* root meristem cell types using scRNA-seq. (**A**) Color-coded tSNE plot (right) showing the classification of 5145 high-quality (UMI count >17,290) cells into distinct cell identities corresponding to the schematic representation of the root meristem on the left. Gray dots represent predicted doublet cells. All inferred and validated developmental trajectories are projected onto the tSNE plot as black lines. Cells within the dashed line are initials. QC, quiescent center; ppc, phloem procambium; se, sieve element; cc, companion cell; px, protoxylem; mx, metaxylem. (**B**) Dot plot showing the expression of known tissue-specific reporter genes in the scRNA-seq dataset, validating the annotation of tissue-specific clusters. The size of the circles represents the percentage of cells with expression (pct. exp.), whereas the color indicates the scaled average expression (avg. exp. scale). Dashed boxes represent major tissue types and cellular stages present in the root. (**C**) Projection of predicted ploidy levels of each cell onto the tSNE plot.

which undergoes continuous divisions as it passes through the meristem, no correlation was found. Cells in the quiescent center and hair cell (trichoblast) clusters mostly contained 2C and 16C cells, respectively (Fig. 1C).

Both newly generated reporter lines and ploidy analysis confirm the inferred developmental trajectories of all main cell identities. Our results thus allow the identification of distinct subclusters linked to the developmental stage of each cell identity, establishing a developmental blueprint for all root cell lineages, including progenitor populations for several cell identities.

TMO5/LHW targets are enriched in root hair cells

We next intersected our scRNA-seq root dataset with the set of 273 genes identified by bulk transcriptome analysis to be target genes of the TMO5/LHW complex (*11*). About 80 target genes (29%) were predicted to be expressed in root hair (trichoblast) cells, of which 47 (17%) were expressed only in trichoblast cells (Fig. 2A and fig. S13). Such expression patterns were not expected given the literature on TMO5/LHW function in vascular proliferation (*1*, *2*, *5*, *6*) and the overlapping expression domain in the young xylem cells (*2*). We confirmed the induction and expression pattern of a subset of these genes by quantitative real-time fluorescence polymerase chain reaction (QPCR) and promoter–green fluorescent protein (GFP) fusions, respectively (Fig. 2B and fig. S14A). The trichoblast-specific expression patterns of these target genes thus suggest a putative role for TMO5/LHW in the regulation of root hair development or patterning. Although homozygous *tmo5* single and *tmo5 tmo5-like1* double mutants showed normal root hair densities under standard growth conditions (high phosphate), misexpression of *TMO5* and *LHW* in all cells of the root meristem (p*RPS5A*::TMO5-GR x p*RPS5A*::LHW-GR or dGR) resulted in a strong increase in root hair density (Fig. 2, C and D), whereas misexpression of unrelated basic helix-loop-helix (bHLH) factors did not result in this root hair density increase (fig. S15). A strong increase in root hair density can also be observed in wild-type roots grown under phosphate-limiting conditions [Fig. 2, C and D, and table S2; see fig. S16 and supplementary materials for a detailed description of three-dimensional (3D) root hair quantifications] (*22*–*24*). Auxin biosynthesis, transport, and signaling are all required for this root hair response to phosphate-limiting conditions (*22*): Under low-phosphate conditions, auxin signaling is induced in the columella and lateral root cap region and in xylem cells, where the TMO5/LHW dimer is active (*2*, *4*). Auxin-dependent TMO5 function is required for the root hair response to low phosphate, because *tmo5 tmo5-like1* double mutants were less sensitive to these limiting conditions (Fig. 2, C and D, and table S2). Phosphate starvation genes (*25*) were, however, still induced in this mutant background (fig. S14B), suggesting that perception was unaffected. Moreover, induction of several root hair–specific TMO5/LHW target genes was *tmo5 tmo5-like1*–dependent (fig. S14C). The TMO5 homologs seem redundantly required for this response, because we found no significant difference in root hair density between wild type and the *tmo5* single mutant in low-phosphate conditions (Fig. 2, C and D) (*26*). Taken together, these results show that the increase in root hair density under low-phosphate conditions specifically requires TMO5 activity. Because low-phosphate conditions have also been associated with changes in root hair length (*27*), we quantified this parameter in our mutant lines and treatments and found similar responses (fig. S14D; see fig. S16 and supplementary materials for details

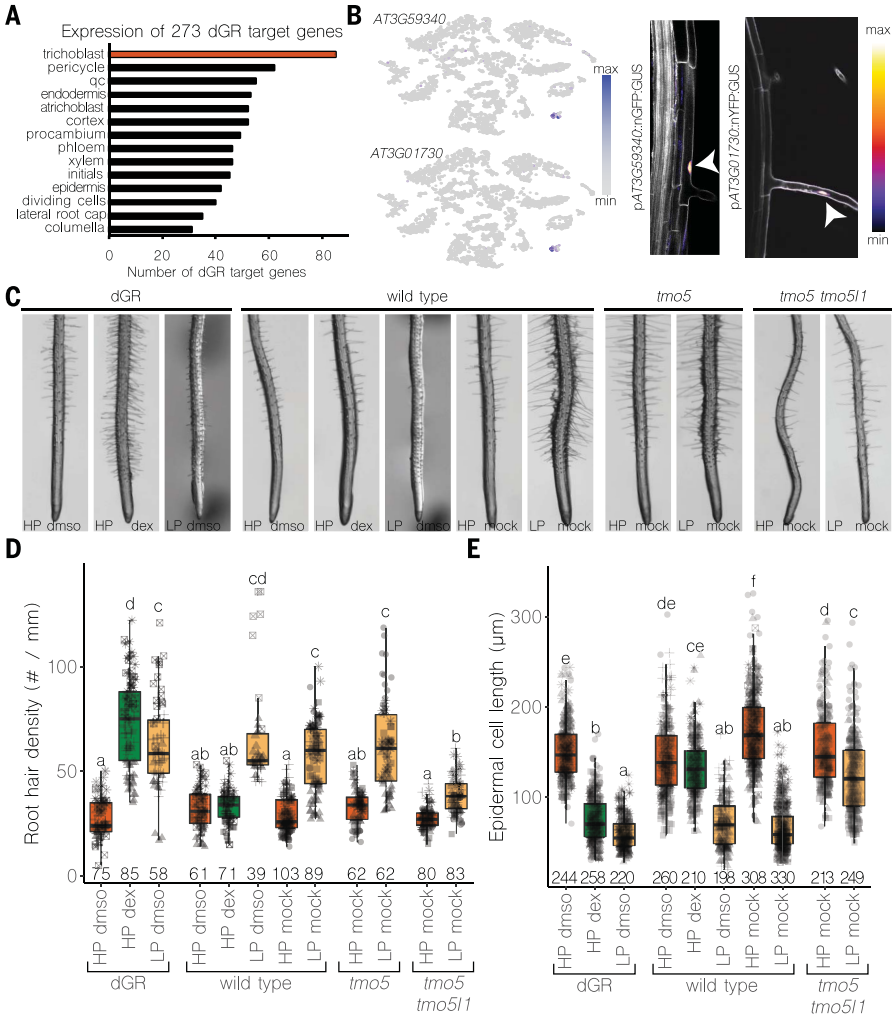


Fig. 2. TMO5 activity is required for root hair responses to low-phosphate conditions. (**A**) Number of TMO5/LHW target genes expressed in each of the tissue types of the *Arabidopsis* root meristem. Note the high number of trichoblast-expressed genes. (**B**) Predicted (left) and validated (right) expression of root hair–specific target genes in the root hair. Arrowheads indicate nuclear expression. Images on the right are confocal laser scanning micrographs. (**C**) Stereo microscope images showing root hair phenotype of dGR [induced or noninduced with dexamethasone (dex)] and wild type, *tmo5* single mutant, and *tmo5 tmo5like1* double mutants grown under control conditions (high phosphate, HP) or phosphate-limiting conditions (low phosphate, LP). (**D** and **E**) Quantification of the root hair density (D) and epidermal cell length (E) of the lines depicted in (C). "Mock" indicates that the growth media contained no additives. Lowercase letters on top of box plots indicate significantly different groups as determined by one-way analysis of variance (ANOVA) with post-hoc Tukey honest significant difference (HSD) testing ($p < 0.001$); the number of individuals is shown at the bottom of the plot, and biological repeats are indicated using different symbols. Boxes represent the 1st and 3rd quartiles, and the center line represents the median. dmso, dimethyl sulfoxide.

on quantification). Thus, activity of the TMO5/LHW complex is required for the complete root hair response to low-phosphate conditions.

To understand the cellular basis of the root hair density increase in response to low-phosphate conditions, we quantified several parameters that could contribute to this effect (including root length, meristem length, and epidermal cell length; Fig. 2E and figs. S14, E and F, and S16) in wild-type and *tmo5 tmo5-like1* roots. Although the *tmo5 tmo5-like1* mutant showed reduced root and meristem length in control conditions compared with wild type, this did not result in altered root hair densities (Fig. 2D and fig. S14, E and F). This suggests that root and meristem length are not directly contributing to changes in root hair density. Fitting with the observed changes in root hair density, epidermal cell length was not different in wild type and *tmo5 tmo5-like1* under control conditions but was decreased in wild type and was significantly less decreased in the *tmo5 tmo5-like1* double mutant under low-phosphate conditions (Fig. 2E). Similar results were obtained upon dGR induction (Fig. 2E). These results suggest that a reduction of epidermal cell length in the root hair zone drives the low-phosphate response, leading to an increase in root hair density.

TMO5/LHW-dependent cytokinin controls root hair responses to low phosphate

To understand how TMO5/LHW in the xylem might be involved in the low-phosphate response of root hairs in the epidermis, we analyzed expression of the transcriptional p*TMO5*::n3GFP reporter line (*28*). *TMO5* expression increased in response to low-phosphate conditions (*29*) (Fig. 3, A and B), consistent with the reported increase in auxin signaling under these limiting conditions (*22*) and the auxin-inducibility of *TMO5* (*2*). No ectopic expression of *TMO5* was observed in the trichoblast cells (Fig. 3A). Furthermore, increasing TMO5 levels only in the xylem axis or in the vascular bundle [using p*TMO5*::TMO5:GR (*1*) or using newly generated p*SHR*::TMO5:GR and p*WOL*::XVE>>TMO5 lines] was sufficient to increase root hair density (Fig. 3, C to E, and fig. S17, A to C). This suggests that the effect of TMO5 on the low-phosphate-induced root hair density increase is cell nonautonomous.

The TMO5/LHW complex binds the *LOG4* promoter, thus promoting cytokinin biosynthesis. Cytokinin can then diffuse to neighboring cells, where its perception induces cell proliferation (*1*, *6*). To investigate if low-phosphate conditions might lead to an increased induction of the cytokinin signaling pathway in epidermal cells, we analyzed the pTCSn::ntdTomato cytokinin-signaling reporter (*11*, *30*). Under low-phosphate conditions, the TCSn reporter was induced in the epidermal cells of the root meristem (Fig. 4A). Additionally, A-type *ARABIDOPSIS* RESPONSE REGULATORS (ARRs) (including *ARR4*, *5*, *6*, *8*, *9*, *12* and *15*) that our scRNA-seq dataset showed to be expressed in trichoblast cells were up-regulated upon TMO5/LHW induction (*11*) (figs. S13 and S14A). Additionally, expression of three trichoblast-restricted TMO5/LHW target genes was found to be induced by exogenous cytokinin treatment (fig. S18A), fitting with published data (*31*). These results suggest an increased induction of the cytokinin signaling pathway in the epidermis under low-phosphate conditions and fit with the published effect of cytokinin on cell length (*32*). To show that increased cytokinin levels and/or signaling might lead to an increase in root hair density owing to a reduction in epidermal cell lengths, we next treated wild-type roots with 0.1 µM 6-benzylaminopurine (BAP), a synthetic cytokinin. Indeed, treatment with cytokinin under high-phosphate conditions mimicked the root hair density–promoting effect of low-phosphate conditions (Fig. 4B) and a reduction in epidermal cell length (fig.

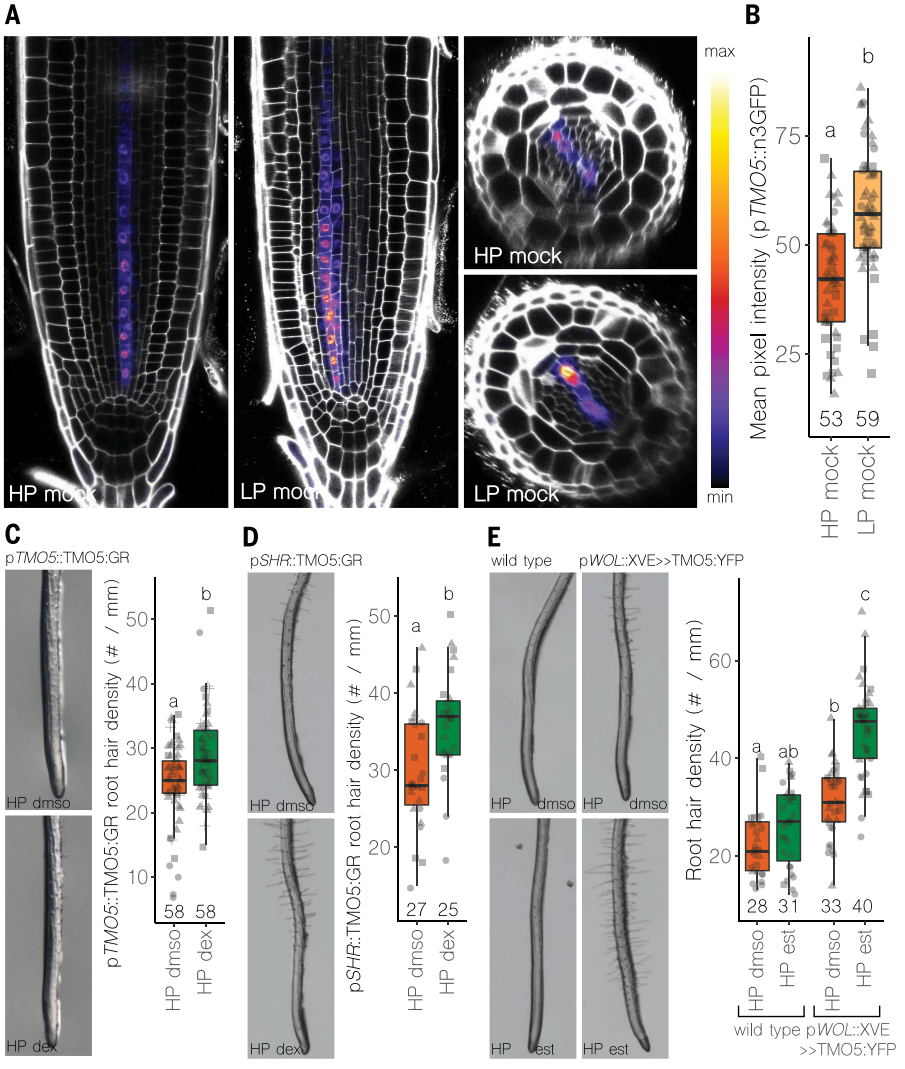


Fig. 3. Vascular TMO5/LHW expression increases root hair density. (**A** and **B**) Expression (A) and quantification (B) of the p*TMO5*::n3GFP reporter line in the root meristem under high- and low-phosphate conditions by confocal microscopy. (**C** to **E**) Root hair phenotype (by stereo microscopy) and quantification of wild-type, p*TMO5*::TMO5:GR, p*SHR*::TMO5:GR, and p*WOL*::XVE>>TMO5:YFP roots grown under high-phosphate conditions or induced by dexamethasone or estradiol (est) (see Fig. 2 for wild-type control). Lowercase letters on top of box plots indicate significantly different groups as determined by one-way ANOVA with post-hoc Tukey HSD testing [$p < 0.001$ in (C) and (E), and $p < 0.01$ in (D)]; the number of individuals is shown at the bottom of the plot, and biological repeats are indicated using different symbols. Boxes represent the 1st and 3rd quartiles, and the center line represents the median.

S18B). Additionally, cytokinin was sufficient to rescue the root hair density and epidermal cell length effects to low-phosphate–like responses in the *tmo5 tmo5-like1* double mutant (Fig. 4B and fig. S18B). To investigate the possible role of ethylene in this response (*33*), we next analyzed the responses of the *ein3 eil1* double mutant in downstream ethylene signaling on low-phosphate medium and upon cytokinin treatment and found no difference compared with wild-type plants (fig. S19A), fitting with published reports that cytokinin effects on root hair length are not dependent on ethylene signaling (*34*) and the fact that ethylene response markers (*33*) were not uniformly altered upon TMO5/LHW induction (fig. S19B). Given that the upstream *ein2* receptor single mutant does show some resistance to cytokinin but not low-phosphate treatment (fig. S19A), we cannot rule out involvement of complex cytokinin-ethylene hormonal crosstalk during this developmental process.

To provide additional genetic support for the hypothesis that cytokinin signaling in trichoblast cells drives the root hair response to low-phosphate conditions in a TMO5-dependent manner, we first examined the effect of reduced cytokinin levels by analyzing the *log347* triple biosynthesis mutant (*8*) and by analyzing a newly generated p*RPS5A*::CKX3 transgenic line that expresses increased levels of the CKX3 cytokinin conjugating enzyme (*35*). Both genetic tools to reduce cytokinin levels resulted in an inhibition of the low-phosphate response (Fig. 4C and fig. S18C). To strengthen the hypothesis that vascular-derived cytokinin is responsible for the root hair response to low-phosphate conditions, we generated a vascular-specific estradiol-inducible *LOG4* transgenic line (p*WOL*::XVE>>LOG4). Upon induction of the cytokinin biosynthetic gene *LOG4* only in the vascular domain, an increase in root hair density was observed (Fig. 4D). Additionally, we complemented the *log 1234578* heptuple mutant (*9*), which has very low levels of active cytokinin, with *LOG4* expressed only in the TMO5 domain (*1*). Under phosphate-limiting conditions, p*TMO5*::LOG4 expression was sufficient to restore a wild type–like response (fig. S20). Taken together, these experiments show that vascular-derived cytokinin is capable of triggering responses in the trichoblast cells. Thus, vascular-derived cytokinin can drive the root hair response to low-phosphate conditions in a TMO5-dependent manner.

Previously, prolonged low-phosphate conditions were shown to increase the number of cortex cell files and modify epidermal cell fates (*36*). To understand if these parameters might contribute to our observed increase in root hair density, we first analyzed the number of cells in radial sections of cortex and epidermal cell files because these determine the number of root hairs (*36*). Although prolonged growth under phosphate-deprived conditions leads to an increase in the number of cortex cells (*36*) (fig. S21), this was not observed after 10 days, a time point used in all of our experiments (table S2), suggesting that the effects observed in our experiments are not due to additional cortex cells. We next examined the possible change in cell fate by examination of epidermal cell identity using markers for hair (p*COBL9*::GFP) (*37*) and nonhair (p*GL2*::GFP) (*38*) cell files. Epidermal cell identities were mixed upon cytokinin treatment, similar to the effect observed under low-phosphate conditions (Fig. 5, A to C) (*36*). This cytokinin-dependent effect most likely feeds into the known pathways determining epidermal cell identity, because cytokinin treatment was not able to induce hair formation in the *cpc try* double mutant (fig. S22) (*39*). Moreover, significantly more root hairs were formed in nonhair positions upon dGR induction and exogenous cytokinin treatment and under low-phosphate conditions (Fig. 5D). This low-phosphate effect was absent in the *tmo5 tmo5-like1* mutant and in plants with reduced cytokinin signaling levels (Fig. 5, D and E). These results suggest that alterations in the epidermal cell identity contribute to the observed increase in root hair density under low-phosphate conditions.

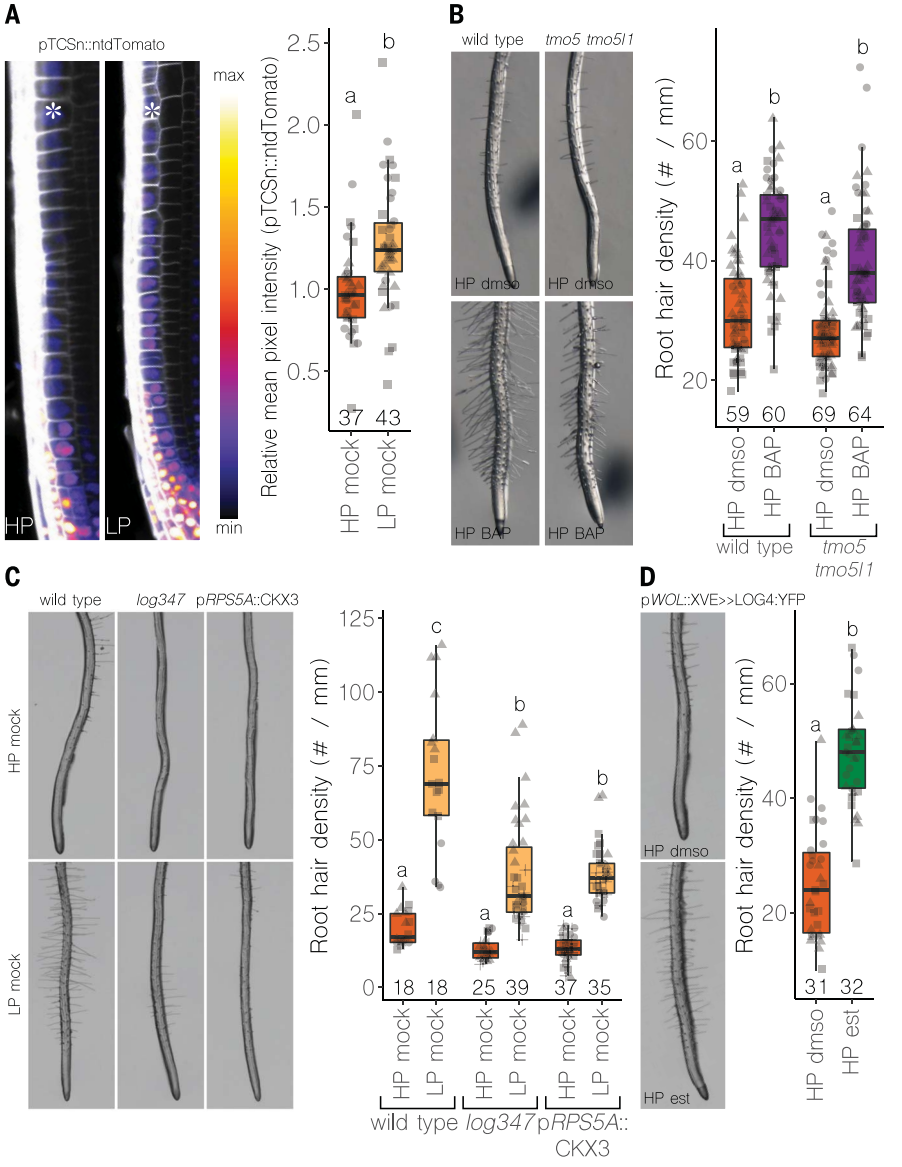


Fig. 4. TMO5/LHW-dependent cytokinin triggers root hair responses. (**A**) Expression and quantification of pTCSn::ntdTomato in the root meristem under high- and low-phosphate conditions by confocal microscopy. Asterisks indicate epidermal cell layer. (**B**) Root hair phenotype (by stereo microscopy) and quantification of wild-type and *tmo5 tmo5like1* roots grown under high-phosphate conditions or induced by cytokinin (BAP). (**C**) Root hair phenotype (by stereo microscopy) and quantification of wild-type, *log347*, and p*RPS5A*::CKX3 roots grown under high- or low-phosphate conditions. (**D**) Root hair phenotype (by stereo microscopy) and quantification of p*WOL*::XVE>>LOG4:YFP roots grown under high-phosphate conditions or induced by estradiol (see Fig. 3 for wild-type control). Lowercase letters on top of box plots indicate significantly different groups as determined by one-way ANOVA with post-hoc Tukey HSD testing ($p < 0.001$); the number of individuals is shown at the bottom of the plot, and biological repeats are indicated using different symbols. Boxes represent the 1st and 3rd quartiles, and the center line represents the median.

Outlook

We show that the vascular bHLH heterodimer TMO5/LHW controls root hair density by modifying epidermal cell length and cell fates. Phosphate deficit may trigger increased auxin signaling in xylem cells, inducing the TMO5/LHW pathway and downstream local cytokinin biosynthesis. Cytokinin may then diffuse outward to direct the length and fates of outer trichoblast cells. As such, this hormone signaling cascade spans multiple tissue layers in the meristem to regulate roots that are foraging for phosphate.

Materials and Methods

Plant material and growth conditions

All seeds were surface-sterilized, sown on solid ½ Murashige and Skoog (MS) plates without sucrose, and stratified at 4°C for 2 days before they were grown at 22°C in continuous light conditions. Ten-day-old seedlings were transferred to soil and grown in greenhouse conditions. For phosphate-limiting experiments, plants were grown for 10 days on plates containing ½ modified MS salts (M407, Phyto Technology Laboratories), 0.55 mM myo-inositol, 20.6 mM NH_4NO_3, 18.8 mM KNO_3, 0.8% washed agar, and 1.25 mM or 100 μM KH_2PO_4 [for high-phosphate (HP) or low-phosphate (LP), respectively]. Dexamethasone (dex) treatment was performed by either germinating seeds on 10 μM dex–supplemented medium or by transferring plants from ½ MS to 10 μM dex–supplemented medium and continuing growth for the indicated time. BAP and β-estradiol (est) treatments were performed by germination on plates supplemented with 0.1 μM BAP or 10 μM est, respectively. BAP, dex, and est were dissolved in dimethyl sulfoxide as solvent (DMSO, 10 mM stock solutions each). Addition of DMSO to the medium was found to have an effect on root hair length, largely canceling out the effect of limiting phosphate conditions on hair length. This effect was limited to hair length, because a large phosphate-dependent effect could still be observed at the level of root hair density (fig. S23). Media containing no additives is referred to as "mock." The *Arabidopsis thaliana* (L.) Heynh. Col-0 ecotype served as wild-type control in all experiments, unless otherwise stated. Plant lines used in this study: p*RPS5A*::TMO5:GR-p*RPS5A*::LHW:GR or dGR (*11*), p*TMO5*::n3GFP (*28*), pTCSn::ntdTomato (*11*), *tmo5* (*2*), *tmo5 tmo5l1* (*2*), *log3 log4 log7* (*8*), *log 1234578 heptuple* mutant (*9*), p*TMO5*::LOG4 in *log 1234578 heptuple* mutant (*1*), p*COBL9*::GFP (*37*), p*GL2*::GFP (*38*), p*WOL*::erGFP (*40*), *ein2* (*41*), *ein3 eil1* (*42*), p*GL2*::GUS in Ws ecotype (*39*), p*GL2*::GUS in *cpc try* double mutant (*39*), p*35S*::IBH1 (*43*), p*35S*::IBL1 (*43*), p*35S*::MYC2 (*44*), and TRANSPLANTA lines TPT_1.01260.2C, TPT_4.09180.1C, TPT_1.01260.2I, TPT_1.05805.1C, TPT_1.05805.1E, TPT_1.05805.1H, TPT_1.62975.1A, TPT_1.62975.1B, TPT_1.62975.1D, TPT_1.62975.1F, TPT_1.62975.1H, TPT_1.62975.1I, TPT_1.68920.2A, TPT_1.68920.2F, TPT_2.18300.1B, TPT_2.18300.1C, TPT_2.18300.1I, TPT_2.22750.2G, TPT_4.09180.1B, and TPT_1.01260.2G (*45*).

Protoplasting conditions and fluorescence-activated cell sorting

Protoplasting was performed as described previously (*46*) with minor modifications. Briefly, root tips of 6-day-old seedlings were cut and incubated in protoplasting Solution B {1.5% (wt/vol) cellulysin and 0.1% (wt/vol) pectolyase in Solution A [600 mM mannitol, 2 mM $MgCl_2$, 0.1% (wt/vol) bovine serum albumin (BSA), 2 mM $CaCl_2$, 2 mM MES, 10mM KCl, pH 5.5]} for about 1 hour at room temperature. Conditions were optimized to ensure that the

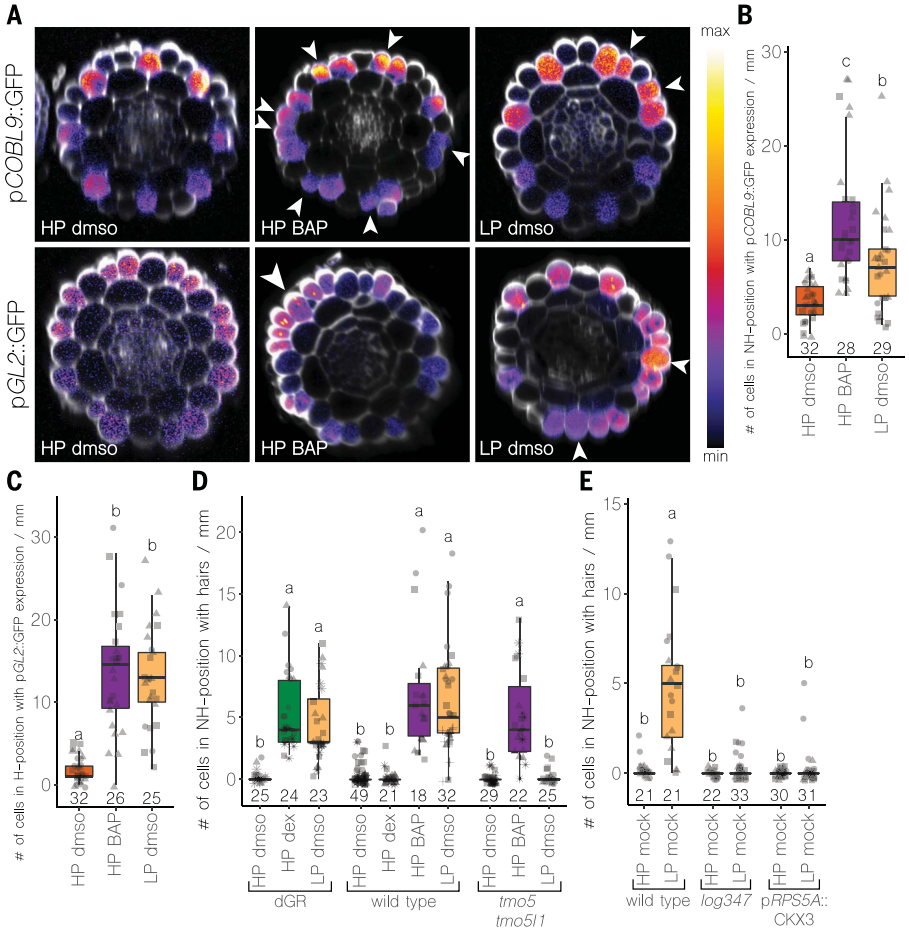


Fig. 5. Cytokinin scrambles epidermal cell identities. (**A**) Expression of p*COBL9*::GFP and p*GL2*::GFP in roots grown under high-phosphate, high-phosphate cytokinin (HP BAP), or low-phosphate conditions. Arrowheads indicate cells with ectopic expression of the marker line. (**B** and **C**) Quantification of number of cells per millimeter of root in nonhair (NH) or hair (H) positions with GFP expression. (**D** and **E**) Quantification of the number of cells in nonhair positions that form root hairs, along 1 mm of root from dGR, wild type, *tmo5*, *tmo5 tmo5l1*, *log347*, or p*RPS5A*::CKX3 grown under the indicated conditions. Lowercase letters on top of the box plots indicate significantly different groups as determined by one-way ANOVA with post-hoc Tukey HSD testing ($p < 0.001$); the number of individuals is shown at the bottom of the plot, and biological repeats are indicated using different symbols. Boxes represent the 1st and 3rd quartiles, and the center line represents the median.

inner vascular cell types were dissociated well (even if this was to some extent at the expense of the outer cell files) by microscopic analysis of the protoplasting at regular intervals. Cells were filtered through a 70-μm cell strainer and spun down at 200*g* for 6 min, resuspended in Solution A, filtered through a 40-μm cell strainer, and stained for viability using 4′,6-diamidino-2-phenylindole (DAPI) at 14 μM final concentration. Cells were sorted on a BD Aria II, and protoplasts not containing the DAPI signal were selected for further analysis.

10X genomics sample preparation, library construction, and sequencing

Sorted cells were centrifuged at 4°C at 400*g* and resuspended in phosphate buffered saline (PBS) with 0.04% BSA to yield an estimated concentration of 1000 cells/μl, before being subdivided to generate three replicates that have gone through the rest of the pipeline as independent samples. Cellular suspensions were loaded on a GemCode Single Cell 3′ Gel Bead and Library Kit (V2 chemistry, 10X Genomics) according to the manufacturer's instructions. Sequencing libraries were loaded on an Illumina HiSeq4000 and sequenced following the recommendations of 10X Genomics at the VIB Nucleomics Core (VIB, Leuven).

Raw data processing and generation of gene expression matrix

Demultiplexing of the raw sequencing data was done by the 10x CellRanger (version 2.0.2) software "cellranger mkfastq," which wraps Illumina's bcl2fastq. The fastq files obtained after demultiplexing were used as the input for "cellranger count," which aligns the reads to the *Arabidopsis thaliana* reference genome (Ensemble TAIR10.40) using STAR and collapses them to UMI counts. The result is a large digital expression matrix with cell barcodes as rows and gene identities as columns. The aggregation of the replicate samples was done using "cellranger aggr." Initial filtering in CellRanger (gene count >200), recovered 15,918 cells. To ensure that we only used high-quality cells for further analysis, we used the filtered data provided by CellRanger. This corresponds to a minimum UMI count of 17,290 in the aggregated data, recovering a final 5145 cells.

Data analysis (clustering, identity assignment, DEGs, and quality control)

All analyses were performed in R (version 3.5.2). Preprocessing of the data was done by the scran (version 1.10.1) and scater (version 1.10.1) packages according to the workflow proposed by the Marioni lab (*47*). Outlier cells were identified based on two metrics (library size and number of expressed genes) and were tagged as outliers when they were three median absolute deviation (MADs) away from the median value of these metrics across all cells. Low-abundance genes were removed using the "calcAverage" function and the proposed workflow by Lun and colleagues (*48*). The raw counts were normalized and log2 transformed by first calculating "size factors" that represent the extent to which counts should be scaled in each library. Detecting highly variable genes, finding clusters, and creating tSNE plots was done using the Seurat pipeline (version 2.3.4). Differential expression analysis for marker gene identification per subpopulation was based on the nonparametric Wilcoxon rank sum test implemented within the Seurat pipeline. Clusters with the same cell annotation based on gene expression analysis were combined to generate a more comprehensible dataset. The relative contribution of each replicate sample to the entire dataset and each individual cluster ranged between 28 and 38%, indicating reproducibility and limited batch effects (fig. S1C). Potential doublets were identified using the DoubletFinder algorithm (version 2.0.0) (*49*). The number of high confidence doublets (indicated in gray on the tSNE plot in Fig. 1A) was below 1% (50 out of 5145 cells). To assess whether a capturing bias of meristem cells was introduced during protoplast isolation, *Arabidopsis* root meristems were imaged using 3D confocal microscopy (see Plant imaging, image processing, and data analysis section below). After segmentation, cells were quantified according to their individual identities, and the in vivo proportions were compared with the proportions within the scRNA-seq dataset. Out of the total number of cells, we

observed a 15% overrepresentation of the inner vascular cell types and a 13% underrepresentation of more outer cell layers (fig. S1D). This was expected, because we optimized the protocol to distinguish inner vascular cells from other cells. No other selective capturing was apparent. The effect of protoplasting on gene expression (*50*) was limited to a maximum 8% of the top 100 DEGs and only observed in the most outer tissue types of the root (fig. S1E)

Data comparison

Comparison of different available *Arabidopsis* root scRNA-seq datasets (*12*–*15*) was done, revealing similarities and differences between these datasets, as shown in Table 1.

Pseudotime trajectory analysis

Pseudotime trajectory analysis was performed with different methods compiled in the Dyno R package (version 0.1.1) (*51*). The optimal method for each cell identity was selected based on prior biological knowledge and the guidelines in the Dyno package. The consensus trajectory was fixed after iterative and reproducible inference calculation. The GNG algorithm was used to infer trajectories for the following cell identities: phloem (seed set at 4), xylem (seed set at 2), cortex (seed set at 1), lateral root cap and columella (seed set at 1), and pericycle (seed set at 13); Slingshot was used for endodermis and cambium (*52*).

For trajectory inference on the xylem subpopulations, cells were ordered along a bifurcating trajectory from xylem initials toward protoxylem and metaxylem (fig. S3A). The identification and validation of the subpopulations and branching points is based on gene expression of known regulators of xylem development such as *ACL5* and *ATHB15* for early metaxylem (*53*, *54*), *TMO5* for early proto- and metaxylem (*2*), *VND2* for intermediate xylem (*55*), and *VND7* and *CESA4* for late protoxylem (*55*, *56*) (fig. S3A). Although our method depends on generating protoplasts and is not efficient at retrieving differentiating cells, we acquired many genes involved in secondary cell wall formation in the protoxylem subcluster but less so in the metaxylem subcluster. All newly generated reporter lines confirmed the predicted pseudotime and tissue-specific expression patterns for the xylem trajectory (fig. S3B).

For the phloem cell lineage, the GNG trajectory algorithm ordered the cells across five cellular states of phloem development with more elaborate branching (fig. S4A). Known reporter genes [including *APL* for non-procambium (*57*), *JUL1* for procambium (*58*), the PEAR gene *DOF5.6* for sieve elements (*10*), and *SUC2* for companion cells (*57*)] confirmed predicted subcluster identities within the phloem developmental trajectories (fig. S4A). Also for this more complex trajectory, all newly generated reporter lines confirmed the predicted spatiotemporal expression patterns (fig. S4B).

A similar analysis and validation were performed on the predicted developmental trajectories and subcluster identities for the procambium, pericycle, endodermis, cortex, epidermis, lateral root cap, and columella clusters (figs. S5 to S11)

Online tool for visualization of data

To allow optimal accessibility of this scRNA-seq dataset by the scientific community, cell type–specific expression patterns of individual genes, expression levels along developmental trajectories, and bulk downloads can be accessed through a freely accessible on-line browser tool (http://bioit3.irc.ugent.be/plant-sc-atlas/), which includes a warning flag for protoplast-induced genes.

Cell ploidy predictions

For predicting the ploidy of individual cells, we correlated the scRNA-seq dataset with a dataset of ploidy-specific gene expression levels in *Arabidopsis* root cortex tissue (*21*). The expression values of 332 previously established ploidy marker genes [supplemental table 5 in (*21*)] were standardized in both datasets separately (zero mean, unit variance), and Pearson correlation coefficients (PCCs) were calculated between the resulting marker gene expression profile in each cell and the marker gene expression profiles of 2C, 4C, 8C, and 16C cells in the cortex dataset. The predicted ploidy level of a given cell was then taken to be the ploidy level for which the PCC between the cortex and cell data was maximal.

Gene selection, cloning, and plant transformation

Genes from the top 20 differentially expressed genes per cell type were selected on the basis of predicted unique expression after manual curation of the feature plots. Transcriptional reporter lines were obtained by amplifying the upstream region of the transcriptional start and cloning this into the pDONRP41R entry vector. The final expression clones were generated by cloning the promoter regions into the pBGWFS7 destination vector using Gateway cloning. Only lines with stable expression patterns in at least 10 independent transformants were kept, resulting in a set of 41 new transcriptional reporter lines. p*TMO5*::TMO5:GR line was generated by first generating a ligation-independent cloning (LIC)–compatible vector containing *TMO5* promoter, after which TMO5:GR was amplified from p*RPS5A*::TMO5:GR template (*1*) and combined using LIC (*59*). The p*RPS5A*::CKX3 line was generated by cloning the coding sequence (CDS) into pPLV28 using LIC. p*SHR*::nYFP was generated by cloning the promoter fragment (*60*) together with nuclear-localized yellow fluorescent protein (nYFP) into pH7m34GW destination vector using Gateway cloning. p*SHR*::TMO5:GR was generated by combining the *SHR* promoter fragment, *TMO5* CDS, and glucocorticoid receptor (GR) tag into pH7m34GW, using Gateway cloning. p*WOL*::XVE>>TMO5:YFP and p*WOL*::XVE>>LOG4:YFP were generated by subcloing the respective CDS with pDONRP4p1r-pWOL-XVE (*60*) and YFP into pH7m34GW using Gateway cloning. All constructs were verified by Sanger sequencing and were transformed into wild type using simplified floral dipping (*61*). All primer sequences used for cloning and sequencing can be found in table S3.

Plant imaging, image processing, and data analysis

Marker lines and different genotypes were either fixed with ClearSee, stained with Calcofluor White and mounted on sides with ClearSee solution (*62*), or directly mounted on slides containing 0.1 mg/ml propidium iodide (PI) for cell wall staining. Confocal microscopy on marker lines was performed on a Leica SP8X using a 40× lens (water-corrected objective lens with NA 1.2). For 3D imaging and segmentation analysis, 6-day-old wild-type roots were stained using the modified pseudo-Schiff–PI (mPS-PI) staining method (*63*). 3D segmentation of the primary root meristem was performed on 3D z-stacks using the MorphoGraphX software (*64*), and cell numbers per cell type were scored in two individual roots (averages used in plot). For root hair analysis, 10-day-old seedling roots were imaged on the plate with a Leica

Table 1. Comparison of different available *Arabidopsis* root scRNA-seq datasets. The "–" indicates data that were not reported.

Dataset	Technology used	Number of cells	Average reads per cell	Median genes per cell
Denyer *et al*., 2019 (*13*)	10X v2	4727	87,000	4276
Jean-Baptiste *et al*., 2019 (*12*)	10X	3121	–	2445
Zhang *et al*., 2019 (*14*)	10X v2	7695	39,667	1875
Ryu *et al*., 2019 (*15*)	10X v2	7522	75,000	5000
This study	10X v2	5145	208,937	6781

stereomicroscope at 32× magnification. Additionally, confocal imaging was applied to generate a 3D stack of a full 1 mm of mature primary root on a Leica SP8X using a 10× dry lens on samples fixed with ClearSee and stained with Calcofluor White. This stack was generated by imaging in the XZY direction, taking 150 images at ~6.7-µm intervals. For a description of the phenotypic parameters that were measured, see fig. S16 and the materials and methods Phenotyping section. Calcofluor White, GFP, YFP, tdTomato, and PI samples were imaged at excitations of 405, 488, 514, 561, and 514 nm, respectively. Calcofluor White, GFP, YFP, tdTomato, and PI were visualized at emissions of 425 to 475 nm, 500 to 535nm, 520 to 550nm, 580 to 630nm, and 600 to 700 nm, respectively. β-Glucuronidase (GUS) staining was performed as described previously (*65*). For aesthetic reasons, some sets of images were globally adjusted in brightness and contrast, rotated, and displayed on a matching background. Plots were generated using R package "ggplot2" (*66*). In all box plots, boxes represent the 1st and 3rd quartiles, and the center line represents the median. Data-point shapes represent biological replicates, and the numbers of measurements are shown above the x axis. Statistical analyses were performed using either one-way analysis of variance (ANOVA) or chi-square tests, and p values below 0.001 were considered significant, unless otherwise stated. All figures were compiled using Adobe Illustrator.

Phenotyping

The approaches used for analysis of the range of phenotypic parameters described in fig. S16 are explained below in more detail. From scanned plates, total primary root lengths were measured using Fiji software. From stereomicroscope images, both the number of root hairs in 1 mm of root and the root hair lengths were measured in the maturation zone of the primary root using Fiji software. From 3D-confocal stacks, the number of cortex and number of epidermal cells (hair versus non-hair position) were counted in one representative cross section. Additionally, the total number of hairs and the number of cells in nonhair position that form root hairs were counted along the full stack. Similar imaging and phenotyping were performed on p*COBL9*::GFP (hair marker) and p*GL2*::GFP (nonhair marker) lines. Here, the number of cells, ectopically expressing the marker, were counted along the entire stack. Meristem length was determined from longitudinal confocal sections of the primary root and measured as the distance from quiescent center to the first elongated cortical cell. Epidermal cell length was determined by measuring the length of 10 trichoblast cells on a longitudinal confocal section from the same 1-mm region of the root where root hairs were counted. A summary of all phenotypic measurements related to cell numbers and root hairs, on respective genotypes and treatments, is presented in table S2.

Relative expression analysis

Total RNA was isolated using an RNeasy kit (QIAGEN). cDNA was synthesized using 1 µg of total RNA and a qScript cDNA SuperMix (QUANTABIO) according to the manufacturer's instructions. Relative expression was measured in triplicate by QPCR on a LightCycler480 II apparatus (Roche) with SYBR GREEN I Master kit (Roche) according to the manufacturer's instructions. Data were analyzed using qBase+ software package (Biogazelle). Expression levels were normalized to those of EEF1a4 and UBC21. All primers used for QPCR analysis can be found in table S3.

REFERENCES AND NOTES

1. B. De Rybel *et al*., Integration of growth and patterning during vascular tissue formation in *Arabidopsis*. *Science* **345**, 1255215 (2014). doi: 10.1126/science.1255215; pmid: 25104393
2. B. De Rybel *et al*., A bHLH complex controls embryonic vascular tissue establishment and indeterminate growth in *Arabidopsis*. *Dev. Cell* **24**, 426–437 (2013). doi: 10.1016/j.devcel.2012.12.013; pmid: 23415953
3. K. Ohashi-Ito, D. C. Bergmann, Regulation of the *Arabidopsis* root vascular initial population by *LONESOME HIGHWAY*. *Development* **134**, 2959–2968 (2007). doi: 10.1242/dev.006296; pmid: 17626058
4. K. Ohashi-Ito, M. Matsukawa, H. Fukuda, An atypical bHLH transcription factor regulates early xylem development downstream of auxin. *Plant Cell Physiol.* **54**, 398–405 (2013). doi: 10.1093/pcp/pct013; pmid: 23359424
5. K. Ohashi-Ito, M. Oguchi, M. Kojima, H. Sakakibara, H. Fukuda, Auxin-associated initiation of vascular cell differentiation by LONESOME HIGHWAY. *Development* **140**, 765–769 (2013). doi: 10.1242/dev.087924; pmid: 23362345
6. K. Ohashi-Ito *et al*., A bHLH complex activates vascular cell division via cytokinin action in root apical meristem. *Curr. Biol.* **24**, 2053–2058 (2014). doi: 10.1016/j.cub.2014.07.050; pmid: 25131670
7. F. Vera-Sirera *et al*., A bHLH-based feedback loop restricts vascular cell proliferation in plants. *Dev. Cell* **35**, 432–443 (2015). doi: 10.1016/j.devcel.2015.10.022; pmid: 26609958
8. T. Kuroha *et al*., Functional analyses of *LONELY GUY* cytokinin-activating enzymes reveal the importance of the direct activation pathway in *Arabidopsis*. *Plant Cell* **21**, 3152–3169 (2009). doi: 10.1105/tpc.109.068676; pmid: 19837870
9. H. Tokunaga *et al*., *Arabidopsis* lonely guy (LOG) multiple mutants reveal a central role of the LOG-dependent pathway in cytokinin activation. *Plant J.* **69**, 355–365 (2012). doi: 10.1111/j.1365-313X.2011.04795.x; pmid: 22059596
10. S. Miyashima *et al*., Mobile PEAR transcription factors integrate positional cues to prime cambial growth. *Nature* **565**, 490–494 (2019). doi: 10.1038/s41586-018-0839-y; pmid: 30626969
11. W. Smet *et al*., DOF2.1 controls cytokinin-dependent vascular cell proliferation downstream of TMO5/LHW. *Curr. Biol.* **29**, 520–529.e6 (2019). doi: 10.1016/j.cub.2018.12.041; pmid: 30686737
12. K. Jean-Baptiste *et al*., Dynamics of gene expression in single root cells of *Arabidopsis thaliana*. *Plant Cell* **31**, 993–1011 (2019). doi: 10.1105/tpc.18.00785; pmid: 30923229
13. T. Denyer *et al*., Spatiotemporal developmental trajectories in the *Arabidopsis* root revealed using high-throughput single-cell RNA sequencing. *Dev. Cell* **48**, 840–852.e5 (2019). doi: 10.1016/j.devcel.2019.02.022; pmid: 30913408
14. T. Q. Zhang, Z. G. Xu, G. D. Shang, J. W. Wang, A single-cell RNA sequencing profiles the developmental landscape of *Arabidopsis* root. *Mol. Plant* **12**, 648–660 (2019). doi: 10.1016/j.molp.2019.04.004; pmid: 31004836
15. K. H. Ryu, L. Huang, H. M. Kang, J. Schiefelbein, Single-cell RNA sequencing resolves molecular relationships among individual plant cells. *Plant Physiol.* **179**, 1444–1456 (2019). doi: 10.1104/pp.18.01482; pmid: 30718350
16. C. N. Shulse *et al*., High-throughput single-cell transcriptome profiling of plant cell types. *Cell Rep.* **27**, 2241–2247.e4 (2019). doi: 10.1016/j.celrep.2019.04.054; pmid: 31091459
17. S. Li, M. Yamada, X. Han, U. Ohler, P. N. Benfey, High-resolution expression map of the *Arabidopsis* root reveals alternative splicing and lincRNA regulation. *Dev. Cell* **39**, 508–522 (2016). doi: 10.1016/j.devcel.2016.10.012; pmid: 27840108
18. R. Ruonala, D. Ko, Y. Helariutta, Genetic networks in plant vascular development. *Annu. Rev. Genet.* **51**, 335–359 (2017). doi: 10.1146/annurev-genet-120116-024525; pmid: 28892639
19. A. Rodriguez-Villalon *et al*., Molecular genetic framework for protophloem formation. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 11551–11556 (2014). doi: 10.1073/pnas.1407337111; pmid: 25049386
20. A. P. Mähönen *et al*., A novel two-component hybrid molecule regulates vascular morphogenesis of the *Arabidopsis* root. *Genes Dev.* **14**, 2938–2943 (2000). doi: 10.1101/gad.189200; pmid: 11114883
21. R. Bhosale *et al*., A spatiotemporal DNA endoploidy map of the *Arabidopsis* root reveals roles for the endocycle in root development and stress adaptation. *Plant Cell* **30**, 2330–2351 (2018). doi: 10.1105/tpc.17.00983; pmid: 30115738
22. R. Bhosale *et al*., A mechanistic framework for auxin dependent *Arabidopsis* root hair elongation to low external phosphate. *Nat. Commun.* **9**, 1409 (2018). doi: 10.1038/s41467-018-03851-3; pmid: 29651114
23. J. E. Salazar-Henao, I. C. Vélez-Bermúdez, W. Schmidt, The regulation and plasticity of root hair patterning and morphogenesis. *Development* **143**, 1848–1858 (2016). doi: 10.1242/dev.132845; pmid: 27246711
24. M. Grebe *et al*., Cell polarity signaling in *Arabidopsis* involves a BFA-sensitive auxin influx pathway. *Curr. Biol.* **12**, 329–334 (2002). doi: 10.1016/S0960-9822(02)00654-1; pmid: 11864575
25. H. Rouached, A. B. Arpat, Y. Poirier, Regulation of phosphate starvation responses in plants: Signaling players and cross-talks. *Mol. Plant* **3**, 288–299 (2010). doi: 10.1093/mp/ssp120; pmid: 20142416
26. Z. H. Chen, G. A. Nimmo, G. I. Jenkins, H. G. Nimmo, BHLH32 modulates several biochemical and morphological processes that respond to P_i starvation in *Arabidopsis*. *Biochem. J.* **405**, 191–198 (2007). doi: 10.1042/BJ20070102; pmid: 17376028
27. Z. Ma, D. G. Bielenberg, K. M. Brown, J. P. Lynch, Regulation of root hair density by phosphorus availability in *Arabidopsis thaliana*. *Plant Cell Environ.* **24**, 459–467 (2001). doi: 10.1046/j.1365-3040.2001.00695.x
28. A. Schlereth *et al*., MONOPTEROS controls embryonic root initiation by regulating a mobile transcription factor. *Nature* **464**, 913–916 (2010). doi: 10.1038/nature08836; pmid: 20220754
29. P. Wu *et al*., Phosphate starvation triggers distinct alterations of genome expression in *Arabidopsis* roots and leaves. *Plant Physiol.* **132**, 1260–1271 (2003). doi: 10.1104/pp.103.021022; pmid: 12857808
30. E. Zürcher *et al*., A robust and sensitive synthetic sensor to monitor the transcriptional output of the cytokinin signaling network in planta. *Plant Physiol.* **161**, 1066–1075 (2013). doi: 10.1104/pp.112.211763; pmid: 23355633
31. K. C. Potter, J. Wang, G. E. Schaller, J. J. Kieber, Cytokinin modulates context-dependent chromatin accessibility through the type-B response regulators. *Nat. Plants* **4**, 1102–1111 (2018). doi: 10.1038/s41477-018-0290-y; pmid: 30420712
32. I. H. Street *et al*., Cytokinin acts through the auxin influx carrier AUX1 to regulate cell elongation in the root. *Development* **143**, 3982–3993 (2016). doi: 10.1242/dev.132035; pmid: 27697901
33. L. Song *et al*., The molecular mechanism of ethylene-mediated root hair development induced by phosphate starvation. *PLOS Genet.* **12**, e1006194 (2016). doi: 10.1371/journal.pgen.1006194; pmid: 27427911
34. S. Zhang *et al*., Multiple phytohormones promote root hair elongation by regulating a similar set of genes in the root epidermis in *Arabidopsis*. *J. Exp. Bot.* **67**, 6363–6372 (2016). doi: 10.1093/jxb/erw400; pmid: 27799284
35. T. Schmülling, T. Werner, M. Riefler, E. Krupková, I. Bartrina y Manns, Structure and function of cytokinin oxidase/dehydrogenase genes of maize, rice, *Arabidopsis* and other species. *J. Plant Res.* **116**, 241–252 (2003). doi: 10.1007/s10265-003-0096-4; pmid: 12721786
36. G. Janes *et al*., Cellular patterning of *Arabidopsis* roots under low phosphate conditions. *Front Plant Sci* **9**, 735 (2018). doi: 10.3389/fpls.2018.00735; pmid: 29922313
37. S. M. Brady, S. Song, K. S. Dhugga, J. A. Rafalski, P. N. Benfey, Combining expression and comparative evolutionary analysis. The *COBRA* gene family. *Plant Physiol.* **143**, 172–187 (2007). doi: 10.1104/pp.106.087262; pmid: 17098858

38. Y. Lin, J. Schiefelbein, Embryonic control of epidermal cell patterning in the root and hypocotyl of *Arabidopsis*. *Development* **128**, 3697–3705 (2001). pmid: 11585796
39. M. Simon, M. M. Lee, Y. Lin, L. Gish, J. Schiefelbein, Distinct and overlapping roles of single-repeat MYB genes in root epidermal patterning. *Dev. Biol.* **311**, 566–578 (2007). doi: 10.1016/j.ydbio.2007.09.001; pmid: 17931617
40. M. Bonke, S. Thitamadee, A. P. Mähönen, M. T. Hauser, Y. Helariutta, APL regulates vascular tissue identity in *Arabidopsis*. *Nature* **426**, 181–186 (2003). doi: 10.1038/nature02100; pmid: 14614507
41. J. M. Alonso, T. Hirayama, G. Roman, S. Nourizadeh, J. R. Ecker, EIN2, a bifunctional transducer of ethylene and stress responses in *Arabidopsis*. *Science* **284**, 2148–2152 (1999). doi: 10.1126/science.284.5423.2148; pmid: 10381874
42. T. Potuschak *et al*., EIN3-dependent regulation of plant ethylene hormone signaling by two *Arabidopsis* F box proteins: EBF1 and EBF2. *Cell* **115**, 679–689 (2003). doi: 10.1016/S0092-8674(03)00968-1; pmid: 14675533
43. M. K. Zhiponova *et al*., Helix-loop-helix/basic helix-loop-helix transcription factor network represses cell elongation in *Arabidopsis* through an apparent incoherent feed-forward loop. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2824–2829 (2014). doi: 10.1073/pnas.1400203111; pmid: 24505057
44. J. Goossens, G. Swinnen, R. Vanden Bossche, L. Pauwels, A. Goossens, Change of a conserved amino acid in the MYC2 and MYC3 transcription factors leads to release of JAZ repression and increased activity. *New Phytol.* **206**, 1229–1237 (2015). doi: 10.1111/nph.13398; pmid: 25817565
45. A. Coego *et al*., The TRANSPLANTA collection of *Arabidopsis* lines: A resource for functional analysis of transcription factors based on their conditional overexpression. *Plant J.* **77**, 944–953 (2014). doi: 10.1111/tpj.12443; pmid: 24456507
46. A. S. Iyer-Pascuzzi, P. N. Benfey, Fluorescence-activated cell sorting in plant developmental biology. *Methods Mol. Biol.* **655**, 313–319 (2010). doi: 10.1007/978-1-60761-765-5_21; pmid: 20734270
47. A. T. Lun, K. Bach, J. C. Marioni, Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016). doi: 10.1186/s13059-016-0947-7; pmid: 27122128
48. A. T. Lun, D. J. McCarthy, J. C. Marioni, A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* **5**, 2122 (2016). doi: 10.12688/f1000research.9501.2; pmid: 27909575
49. C. S. McGinnis, L. M. Murrow, Z. J. Gartner, DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337.e4 (2019). doi: 10.1016/j.cels.2019.03.003; pmid: 30954475
50. K. Birnbaum *et al*., A gene expression map of the *Arabidopsis* root. *Science* **302**, 1956–1960 (2003). doi: 10.1126/science.1090022; pmid: 14671301
51. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019). doi: 10.1038/s41587-019-0071-9; pmid: 30936559
52. K. Street *et al*., Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018). doi: 10.1186/s12864-018-4772-0; pmid: 29914354
53. L. Muñiz *et al*., ACAULIS5 controls *Arabidopsis* xylem specification through the prevention of premature cell death. *Development* **135**, 2573–2582 (2008). doi: 10.1242/dev.019349; pmid: 18599510
54. A. Carlsbecker *et al*., Cell signalling by microRNA165/6 directs gene dose-dependent root cell fate. *Nature* **465**, 316–321 (2010). doi: 10.1038/nature08977; pmid: 20410882
55. M. Kubo *et al*., Transcription switches for protoxylem and metaxylem vessel formation. *Genes Dev.* **19**, 1855–1860 (2005). doi: 10.1101/gad.1331305; pmid: 16103214
56. M. Taylor-Teeples *et al*., An *Arabidopsis* gene regulatory network for secondary cell wall synthesis. *Nature* **517**, 571–575 (2015). doi: 10.1038/nature14099; pmid: 25533953
57. S. Otero, Y. Helariutta, Companion cells: A diamond in the rough. *J. Exp. Bot.* **68**, 71–78 (2017). doi: 10.1093/jxb/erw392; pmid: 27811001
58. V. López-Salmerón, H. Cho, N. Tonn, T. Greb, The phloem as a mediator of plant growth plasticity. *Curr. Biol.* **29**, R173–R181 (2019). doi: 10.1016/j.cub.2019.01.015; pmid: 30836090
59. J. R. Wendrich, C. Y. Liao, W. A. van den Berg, B. De Rybel, D. Weijers, Ligation-independent cloning for plant research. *Methods Mol. Biol.* **1284**, 421–431 (2015). doi: 10.1007/978-1-4939-2444-8_21; pmid: 25757785
60. R. Siligato *et al*., MultiSite gateway-compatible cell type-specific gene-inducible system for plants. *Plant Physiol.* **170**, 627–641 (2016). doi: 10.1104/pp.15.01246; pmid: 26644504
61. S. J. Clough, A. F. Bent, Floral dip: A simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998). doi: 10.1046/j.1365-313x.1998.00343.x; pmid: 10069079
62. R. Ursache, T. G. Andersen, P. Marhavý, N. Geldner, A protocol for combining fluorescent proteins with histological stains for diverse cell wall components. *Plant J.* **93**, 399–412 (2018). doi: 10.1111/tpj.13784; pmid: 29171896
63. E. Truernit *et al*., High-resolution whole-mount imaging of three-dimensional tissue organization and gene expression enables the study of phloem development and structure in *Arabidopsis*. *Plant Cell* **20**, 1494–1503 (2008). doi: 10.1105/tpc.107.056069; pmid: 18523061
64. P. Barbier de Reuille *et al*., MorphoGraphX: A platform for quantifying morphogenesis in 4D. *eLife* **4**, e05864 (2015). doi: 10.7554/eLife.05864; pmid: 25946108
65. B. De Rybel *et al*., A novel aux/IAA28 signaling cascade activates GATA23-dependent specification of lateral root founder cell identity. *Curr. Biol.* **20**, 1697–1706 (2010). doi: 10.1016/j.cub.2010.09.007; pmid: 20888232
66. H. Wickham, *ggplot2, Elegant Graphics for Data Analysis (Use R!)* (Springer, 2009), vol. 1, pp. VIII, 213.

ACKNOWLEDGMENTS

We thank V. Storme for help with statistical analyses and D. Weijers for the use of unpublished materials. **Funding:** This work was supported by funding from the European Research Council (ERC Starting Grant TORPEDO; 714055); the Research Foundation – Flanders (FWO; Odysseus II GOD0515N); Ghent University (BOF20/GOA/012 and BOF18/PDO/151); and a Marie Curie Fellowship (IEF-2009-252503). **Author contributions:** B.D.R., Y.S., and G.V.I. conceived the project; B.D.R., J.R.W., and B.Y. designed experiments; J.R.W., J.V.D., B.D.R., and G.V.I. generated samples for scRNA-seq; N.V. performed scRNA-seq; N.V., K.V., and Y.S. analyzed scRNA-seq data and performed trajectory analyses; K.V. generated the online browser tool; W.S., B.W., E.M., H.E.A., J.N., B.Y., and J.R.W. generated reporter lines to validate the dataset; S.M. performed ploidy analysis; B.Y., C.V.d.V., M.M., and J.R.W. analyzed effects on root hair growth; B.D.R. and Y.S. supervised the project; and B.D.R. wrote the paper with input from all authors. **Competing interests:** None declared. **Data and materials availability:** The single-cell data can be accessed through a freely accessible online browser tool (https://bioit3.irc.ugent.be/plant-sc-atlas/), and raw data can be accessed at NCBI with GEO number GSE141730. All other data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials. Material requests should be directed to the corresponding authors.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/eaay4970/suppl/DC1
Materials and Methods
Figs. S1 to S23
Tables S1 to S3
References
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

24 June 2019; resubmitted 17 July 2020
Accepted 4 September 2020
Published online 17 September 2020
10.1126/science.aay4970

RESEARCH ARTICLE

CORONAVIRUS

Immune life history, vaccination, and the dynamics of SARS-CoV-2 over the next 5 years

Chadi M. Saad-Roy[1]*, Caroline E. Wagner[2,3,4]*, Rachel E. Baker[2,3], Sinead E. Morris[5], Jeremy Farrar[6], Andrea L. Graham[2], Simon A. Levin[2], Michael J. Mina[7], C. Jessica E. Metcalf[2,8], Bryan T. Grenfell[2,8,9]†

The future trajectory of the coronavirus disease 2019 (COVID-19) pandemic hinges on the dynamics of adaptive immunity against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2); however, salient features of the immune response elicited by natural infection or vaccination are still uncertain. We use simple epidemiological models to explore estimates for the magnitude and timing of future COVID-19 cases, given different assumptions regarding the protective efficacy and duration of the adaptive immune response to SARS-CoV-2, as well as its interaction with vaccines and nonpharmaceutical interventions. We find that variations in the immune response to primary SARS-CoV-2 infections and a potential vaccine can lead to markedly different immune landscapes and burdens of critically severe cases, ranging from sustained epidemics to near elimination. Our findings illustrate likely complexities in future COVID-19 dynamics and highlight the importance of immunological characterization beyond the measurement of active infections for adequately projecting the immune landscape generated by SARS-CoV-2 infections.

The novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) betacoronavirus (β-CoV) pandemic has resulted in substantial morbidity and mortality, with over 27 million confirmed cases worldwide at the time of writing. To curb viral transmission, nonpharmaceutical interventions (NPIs), including business and school closures, restrictions on movement, and total lockdowns, have been implemented to various degrees around the world. Major efforts to develop effective vaccines and antivirals are ongoing.

Understanding the future trajectory of this disease requires knowledge of the population-level landscape of immunity, generated by the life histories of SARS-CoV-2 infection or vaccination among individual hosts. We show that the nature of secondary infection, particularly the degree of acquisition, retransmission, and clinical severity of subsequent infections with the same pathogen, is particularly important. The nature of acquired immune responses after natural infection varies substantially among pathogens. At one end of this immune spectrum, natural infection with measles (*1*) or smallpox (*2*) virus results in lifelong protection from the reacquisition and retransmission of secondary infections. Many other infections [e.g., influenza (*3*) and respiratory syncytial virus (RSV) (*4*)] confer imperfect or transient clinical and transmission-blocking immunity by either pathogen evolution or waning immunological memory. Finally, phenomena such as antibody-dependent enhancement (ADE) associated with prior natural infection [e.g., dengue (*5*)] or a vaccine [e.g., RSV (*6*)] could result in more clinically severe secondary infections. Furthermore, the immunity conferred by vaccines may not provide complete protection against reinfection and/or disease (*7*), and this protection may be inferior to that acquired after natural infection (*8*). Nevertheless, imperfect vaccines that reduce both the clinical severity and transmissibility of subsequent infections (if they do occur) can still provide population-level disease protection (*7*, *9*, *10*).

The nature of the immune response after natural SARS-CoV-2 infection remains an area of active investigation (*11*–*18*). Reports from serological population- and individual-level studies demonstrate that detectable antibody levels can wane over the first few months post-infection (*19*), yet recent findings demonstrate robust antibody responses 4 months after infection (*20*). This is broadly consistent with serum antibody levels against the seasonal coronavirus human coronavirus OC43 (HCoV-OC43) [which belongs to the same β-CoV genus as SARS-CoV-2 (*21*)], which wane on the time scale of a few months (*22*) to 1 year (*23*). Such seasonal β-CoVs (which also include HCoV-HKU1) are thought to cause repeated infections throughout life (*24*), although a significant biennial component in their dynamics implies at least some herd protection (*21*, *25*). This genus also contains other viruses that cause severe infections in humans, including Middle East respiratory syndrome and SARS-CoV-1 coronaviruses (*21*). Whereas humoral immunity to SARS-CoV-1 is believed to last up to 2 to 3 years (*26*, *27*), antigen-specific T cells against this virus were found to be detectable for at least 11 years after infection (*28*). Indeed, T cell–mediated responses likely play a central role in controlling SARS-CoV-2 replication and disease (*14*, *15*). Recent evidence of preexisting T cells (*14*, *15*) and antibodies (*29*) capable of cross-reacting with SARS-CoV-2 suggests that immunological memory responses elicited during infection with seasonal coronaviruses may also affect coronavirus disease 2019 (COVID-19) susceptibility and disease risk. Finally, although it is currently unclear whether ADE influences the pathogenesis of SARS-CoV-2, it has been hypothesized that severe COVID-19 cases may arise from the presence of nonneutralizing antibodies from prior coronavirus infections (*30*), in agreement with earlier proposals for related coronaviruses (*31*–*33*).

Various epidemiological models have been developed to capture how the diversity or variation in immune responses influences population-level infection dynamics. For instance, the well-known Susceptible-Infected-Recovered (SIR) model is suitable for modeling the dynamics of perfectly immunizing infections such as measles (*34*), whereas the Susceptible-Infected-Recovered-Susceptible (SIRS) model captures the epidemiology of imperfectly immunizing infections such as influenza; here, individuals eventually return to a fully or substantially susceptible class after a finite period of immunity, because of either waning memory or pathogen evolution (*35*). More complex compartmental models have also been developed to study infections characterized by intermediate immune responses lying between these two extremes, such as rotavirus (*36*) and RSV (*4*).

Here, we adopt a generalization of these models, the SIR(S) model (*35*), outlined schematically in Fig. 1 and fig. S1, to explore how the pandemic trajectory might unfold for different assumptions regarding the nature of the adaptive immune response to SARS-CoV-2 infection. Because different adaptive immune responses may be associated with variations in the proportion of severe secondary cases, we also consider a range of values for this fraction in order to explore the potential future clinical burden of SARS-CoV-2 infections. The model assumes different infection and immune phenotypes, depending on exposure history [see (*37*) for the full mathematical details]. Specifically, it interpolates between the fully immunizing SIR model, when immunity is lifelong, and the imperfectly immunizing SIRS model via the degree of susceptibility to and

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540, USA. [2]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. [3]Princeton Environmental Institute, Princeton University, Princeton, NJ 08544, USA. [4]Department of Bioengineering, McGill University, Montreal, Quebec H3A 0C3, Canada. [5]Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY 10032, USA. [6]Wellcome Trust, London, UK. [7]Departments of Epidemiology and Immunology and Infectious Diseases, Harvard School of Public Health, Boston, MA 02115, USA. [8]Princeton School of Public and International Affairs, Princeton University, Princeton, NJ 08544, USA. [9]Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: grenfell@princeton.edu

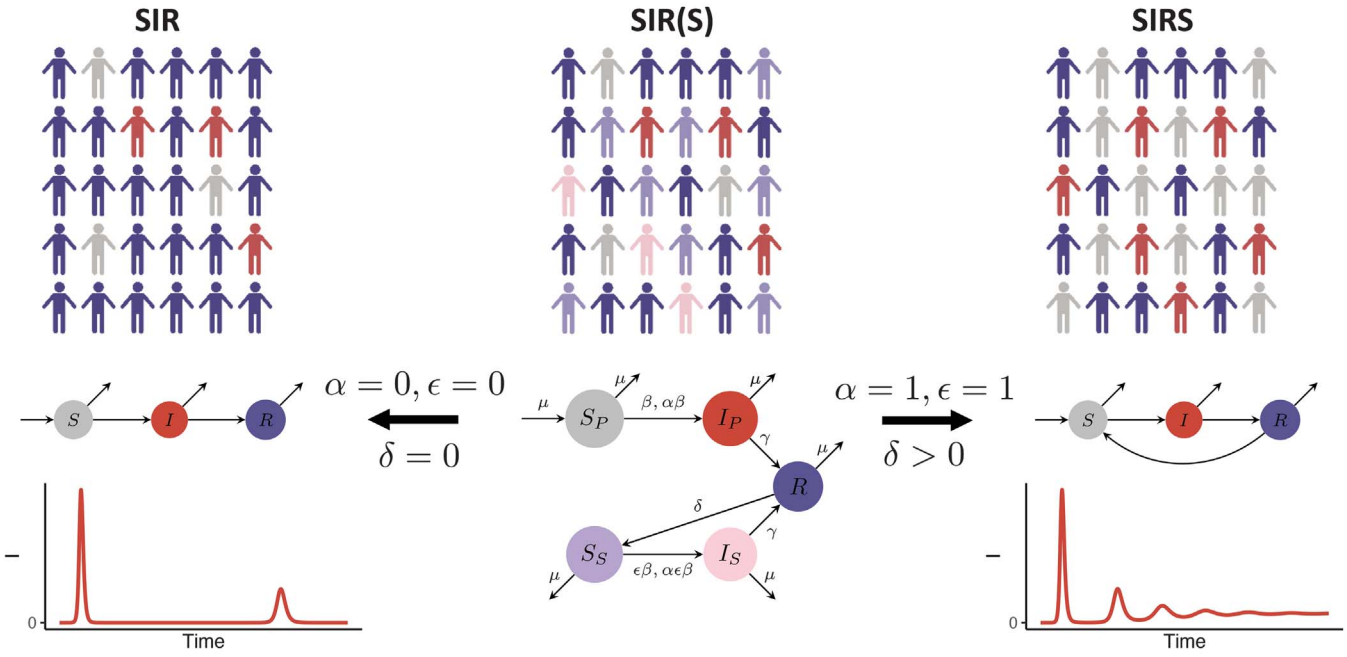


Fig. 1. Schematic of the SIR(S) model with a flowchart depicting flows between immune classes. Here, S_P denotes fully susceptible individuals; I_P denotes individuals with primary infection that transmit at rate β; R denotes fully immune individuals (a result of recovery from either primary or secondary infection); S_S denotes individuals whose immunity has waned at rate δ and are now again susceptible to infection, with relative susceptibility ε; I_S denotes individuals with secondary infection that transmit at a reduced rate αβ; and μ denotes the birth rate (*37*). Illustrations and flowcharts of the limiting SIR and SIRS models are also shown (where individuals are either fully susceptible (*S*), infected (*I*), or fully immune (*R*)), along with a representative time series for the number of infections in each scenario. The population schematics were made through use of (*62*).

transmissibility of secondary infections (quantified by the parameters ε and α, respectively). As shown in the representative time series of Fig. 1, the SIR model results in recurrent epidemics fueled by births following the pandemic peak; by contrast, the SIRS model typically generates shorter interepidemic periods owing to the possibility of reinfection and the buffering of the fully susceptible birth cohort by partially immune individuals (*35*).

We begin by characterizing the effect of temporal changes in the transmission rate brought about by climate and the deployment of NPIs on the predictions of the SIR(S) model under a range of immunity assumptions. Next, we examine the impact of a transmission-reducing vaccine of varying efficacy relative to natural immunity. Finally, we estimate the postpandemic immunity landscape and clinical case burden for different possible futures (*38*) shaped by the various aspects of SARS-CoV-2 biology as well as the presence or absence of these external drivers and interventions, as well as vaccine refusal. To focus on the dynamic impact of natural and vaccinal immunity, we begin with a simple homogeneous model, which averages across known heterogeneities in COVID-19 transmission and severity [age (*39*), superspreading events (*40*), etc.]. We then use heterogeneous model extensions to show that such heterogeneities do not impact our exploration of qualitative medium-term dynamics under different immunological scenarios.

Seasonal transmission rates and the deployment of NPIs

Medium-term dynamics will be shaped by changes in the magnitude of transmission. To explore the effect of NPIs, we considered two different scenarios for timed reductions in the force of infection to 60% of its original value [in agreement with intermediate levels of social distancing in (*21*)]. In Fig. 2, A to C, we show the time courses of primary and secondary infections, assuming single periods of NPI lasting from weeks 16 to 67 (Fig. 2A) or 16 to 55 (Fig. 2B) and two shorter periods during weeks 16 to 55 and weeks 82 to 93 separated by normal interactions (Fig. 2C). We further assume a seasonal transmission rate derived from the climate of New York City (*37*), although in principle this seasonality could also be derived from other nonclimate factors (*25*). The weekly reproduction numbers corresponding to these three scenarios are shown in fig. S2, D to F. Although these reproduction numbers are based on those obtained for the related β-CoV HCoV-HKU1 and are in general lower than those estimated during the early stages of the SARS-CoV-2 pandemic (*41*), they may be more appropriate for considering the longer-term transmission dynamics.

We find that decreases in the susceptibility to secondary infection, ε, can delay secondary peaks (compare individual time courses for different values of ε in Fig. 2, A to C). However, delayed peaks may then be larger, because of susceptible accumulation (through demography or immune waning) and dynamic resonance. These nonmonotonicities in the timing and size of secondary peaks also occur with climate-driven seasonal transmission in the absence of NPIs (*37*), and the trends are qualitatively similar if NPIs are assumed to be relaxed more gradually (fig. S11). Notably, the delay that social distancing may cause in the timing of the secondary peak can also allow for further accumulation of fully susceptible individuals. This is illustrated in the top panels of Fig. 2D, where the average infection rate per infected individual for fully (βS_P; red curve) and partially ($\varepsilon\beta S_S$; green curve) susceptible individuals for the social distancing scenario outlined in Fig. 2C are shown. We contrast a reduction in susceptibility to secondary infection of 50% (ε = 0.5, left panels) with no reduction in susceptibility to secondary infection (ε = 1, right panels). The corresponding fraction of primary (blue) and secondary (purple) cases are presented in the bottom panels. As can be seen, when the secondary peak does occur, the decrease in susceptibility to secondary infection (ε < 1),

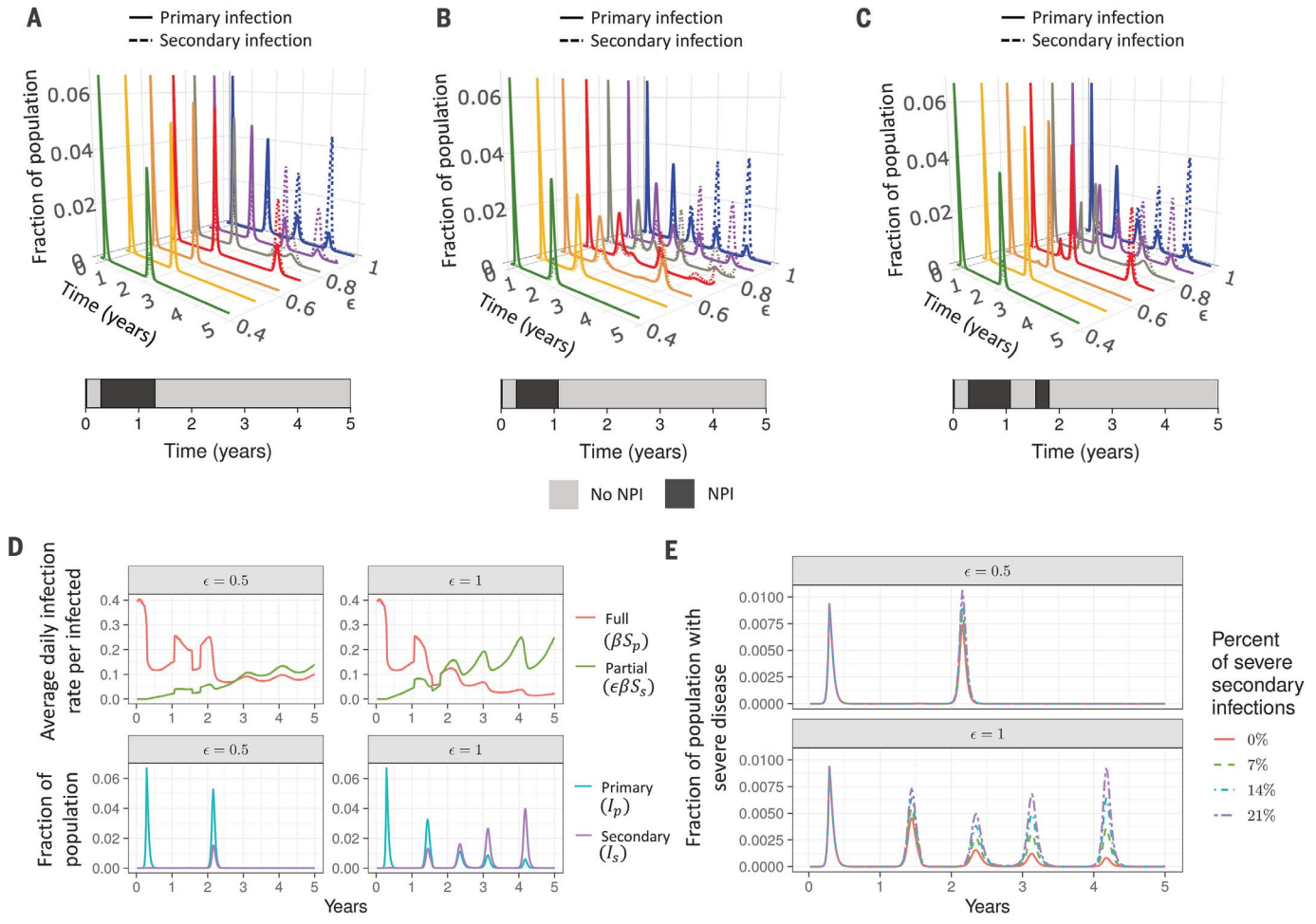


Fig. 2. Seasonality in transmission rates and NPIs modulate disease dynamics. (**A** to **C**) Effect of NPI adoption on the time series of primary (solid lines) and secondary (dashed lines) infections with a seasonal transmission rate derived from the climate of New York City with no lag between seasonality and epidemic onset. NPIs that reduce the transmission rate to 60% of the estimated climate value are assumed to be adopted during weeks 16 to 67 (A), weeks 16 to 55 (B), or weeks 16 to 55 as well as weeks 82 to 93 (C). Colors denote individual time courses for different values of ε. (**D**) Time series of the average daily infection rate per infected individual of fully susceptible (red line) and partially susceptible (green line) individuals (top row) and the fraction of the population that is infected with primary (blue line) and secondary (purple line) infections (bottom row), for ε = 0.5 (left column) and ε = 1 (right column) for the NPI scenario outlined in (C). (**E**) Time series of estimated numbers of severe infections for the NPI scenario defined in (C) for four different estimates of the fraction of severe cases during primary infections ($x_{sev,p}$) and secondary infections ($x_{sev,s}$) with ε = 0.5 (top row) and ε = 1 (bottom row). These are $x_{sev,p}$ = 0.14, $x_{sev,s}$ = 0 (solid red line); $x_{sev,p}$ = 0.14, $x_{sev,s}$ = 0.07 (dashed green line); $x_{sev,p}$ = 0.14, $x_{sev,s}$ = 0.14 (dashed and dotted blue line); and $x_{sev,p}$ = 0.14, $x_{sev,s}$ = 0.21 (purple line with long and short dashes). In all panels, the relative transmissibility of secondary infections and duration of natural immunity are taken to be α = 1 and 1/δ = 1 year, respectively. The effects of NPIs and other parameter variations can be explored interactively at https://grenfelllab.princeton.edu/sarscov2dynamicsplots.

considered in the left panels, results in a greater number of primary infections during the second peak relative to the panels on the right, where ε is 1 [and the secondary infection rate per case (green curves) rises sharply].

Next, an essential part of the planning and management of future SARS-CoV-2 infections is the ability to characterize the magnitude and timing of severe cases requiring hospitalization. In Fig. 2E we consider four possible scenarios for the fraction of severe secondary cases, $x_{sev,s}$ (*37*), on the basis of the scenario depicted in Fig. 2C and assuming 14% of primary cases are severe (*42*): (i) no severe cases associated with secondary infection ($x_{sev,s}$ = 0; solid red line); (ii) a reduced number of severe cases with secondary infection relative to primary infection ($x_{sev,s}$ = 0.07; dashed green line); (iii) comparable proportions of severe cases ($x_{sev,s}$ = 0.14; dashed-dotted blue line); and (iv) a hypothetical greater proportion of severe cases with secondary infection ($x_{sev,s}$ = 0.21; purple line with short and long dashes), possibly owing to phenomena such as ADE. When the assumed fraction of severe subsequent infections is high, the fraction of the population with severe infections during subsequent infection peaks is found to be comparable to or even to slightly exceed that observed during the initial pandemic peak (Fig. 2E). As the proportion of secondary infections increases during the later stages of the pandemic, these findings stress that clinical epidemiological studies of repeat infections will be critical for proper planning of health care systems. We also do not consider any long-term clinical impact of infection here (*43*). The impact of increases in clinical severity with age is addressed below.

Vaccination

The availability of an effective vaccine would be a key intervention against SARS-CoV-2, and numerous candidates are in development (*44*, *45*). Intuitively, if the effective vaccination rate is sufficiently high, then vaccinal

herd immunity generated by a transmission-blocking vaccine could control or eliminate the infection. However, this becomes harder to achieve when vaccinal and natural immunity is imperfect and secondary infections occur, or when logistical or other constraints limit vaccine deployment. We extend the model (*37*) to include a vaccinated class, V, and make the relatively optimistic assumption that a transmission-reducing vaccine begins to be introduced to general populations (the t_{vax}) after 1.5 years. We also consider seasonal transmission rates, as in fig. S3, and the deployment of NPIs according to the scenario described in Fig. 2B. We assume that a constant proportion, ν, ranging from $0\% \le \nu \le 1\%$ of the fully and partially susceptible populations (S_P and S_S), is effectively vaccinated every week and acquires transmission-blocking immunity for, on average, a period $1/\delta_{vax}$. For comparison, it was estimated that in response to the 2009 H1N1 pandemic, one or more doses of the monovalent vaccine were administered to 80.8 million vaccinees during October 2009 to May 2010 in the United States (*46*), which implies a rate of vaccination coverage of about 27% after a period of 8 months for persons aged at least 6 months in the United States, although rates between different nations varied (*47*). This crudely corresponds to a weekly vaccination rate of 1% (*48*) (as with other parameter variations, different scenarios for vaccination can be explored with the accompanying Shiny application). Finally, we assume that the immunity conferred from effective vaccination wanes at rate δ_{vax}, which in general may differ from the waning rate of immunity from natural infection, δ. The modified set of ordinary differential equations in this scenario

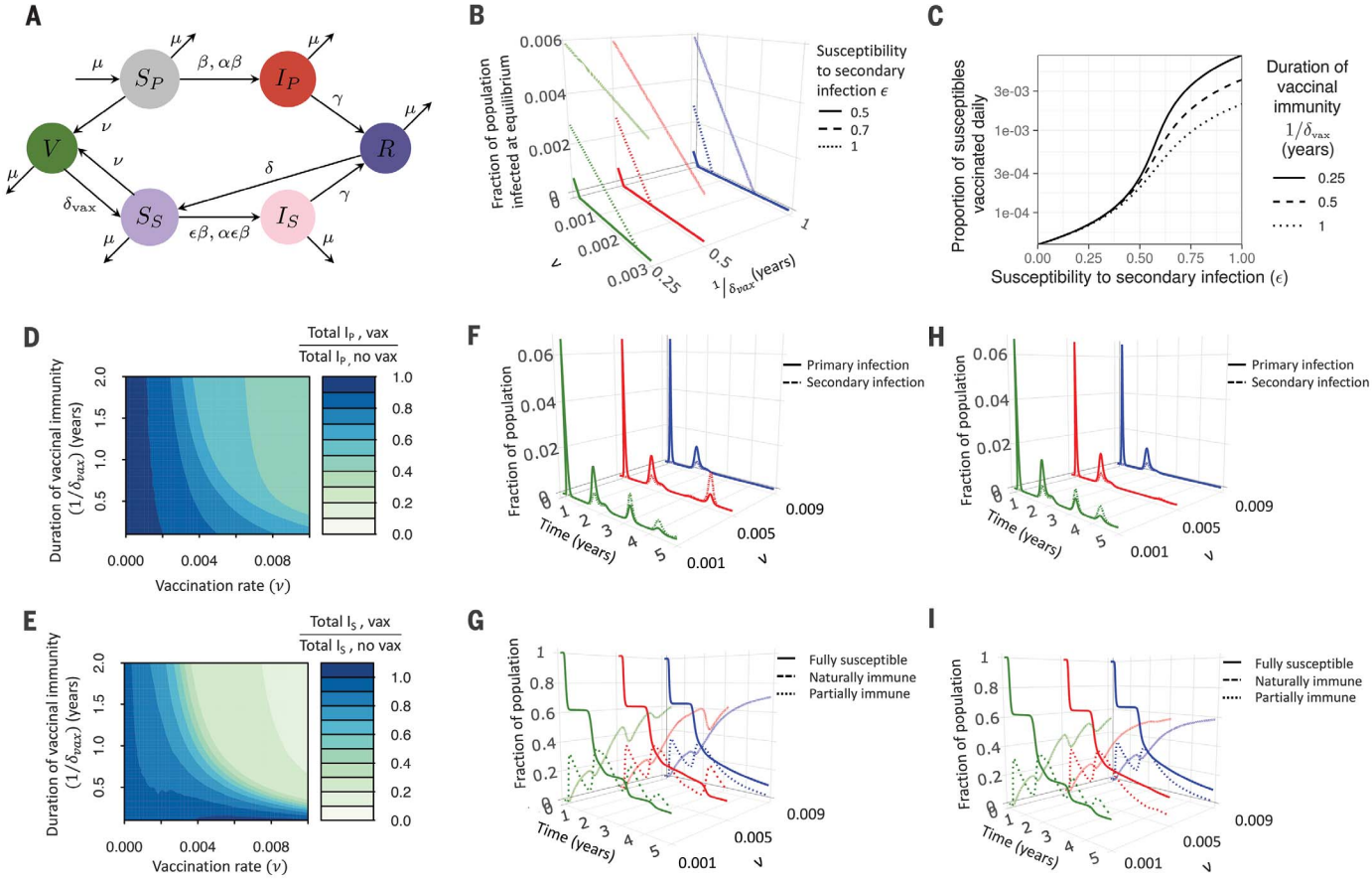


Fig. 3. Impact of vaccination and vaccinal immunity on disease dynamics. (**A**) Modified model flowchart that incorporates a vaccinated class V (*37*). (**B**) Total infected fraction of the population at equilibrium as a function of the vaccination rate ν for different values of the duration of vaccinal immunity ($1/\delta_{vax}$ = 0.25 years, green lines; $1/\delta_{vax}$ = 0.5 years, red lines; and $1/\delta_{vax}$ = 1 year, blue lines) and the susceptibility to secondary infection (ϵ = 0.5, solid lines; ϵ = 0.7, dashed lines; and ϵ = 1, dotted lines). (**C**) Daily proportion of susceptibles who must be vaccinated in order to achieve a disease-free state at equilibrium as a function of ϵ for different values of the duration of vaccinal immunity ($1/\delta_{vax}$ = 0.25 years, solid line; $1/\delta_{vax}$ = 0.5 years, dashed line; and $1/\delta_{vax}$ = 1 year, dotted line). In (B) and (C), the relative transmissibility of secondary infections and duration of natural immunity are taken to be α = 1 and $1/\delta$ = 1 year, respectively, and the transmission rate is derived from the mean value of seasonal New York City–based weekly reproduction numbers ($\bar{R}_0$ = 1.75) (fig. S2C) (*37*). (**D** and **E**) The ratio of the total number of primary (D) and secondary (E) infections with vaccination versus without vaccination, during years 1.5 to 5 (i.e., after the introduction of the vaccine) are plotted as a function of the weekly vaccination rate ν and the duration of vaccinal immunity $1/\delta_{vax}$. (**F** to **I**) Time series of the various immune classes plotted for different values of the vaccination rate ν. The top row [(F) and (H)] contains the time series of primary (I_P, solid lines) and secondary (I_S, dashed lines) infections, whereas the bottom row [(G) and (I)] contains the time series of the fully susceptible (S_P, solid lines), naturally immune (R, dashed lines), and partially immune (S_S, dotted lines) subpopulations. The duration of vaccinal immunity is taken to be $1/\delta_{vax}$ = 0.5 years (shorter than natural immunity) in (F) and (G), and $1/\delta_{vax}$ = 1 year (equal to natural immunity) in (H) and (I). In (D) to (I), the relative susceptibility to secondary infection, relative transmissibility of secondary infections, and duration of natural immunity are taken to be ϵ = 0.7, α = 1, and $1/\delta$ = 1 year, respectively. Vaccination is introduced 1.5 years after the onset of the epidemic (i.e., during the 79th week) following a 40-week period of social distancing during which the force of infection was reduced to 60% of its original value during weeks 16 to 55 (i.e., the scenario described in Fig. 2B), and a seasonal transmission rate derived from the climate of New York City with no lag is assumed.

corresponding to the flowchart in Fig. 3A is also presented in (*37*).

In Fig. 3B, we begin by considering the long-term equilibrium infection burden (*37*) driven by vaccination at a weekly rate ν, for a variety of immunity assumptions. As expected, a reduction in the susceptibility to secondary infections (ε) results in a smaller number of infections at steady state in the absence of vaccination. Further, both ε and the duration of vaccinal immunity ($1/\delta_{vax}$) affect the vaccination rate required to achieve a disease-free state at equilibrium. At the limit of fully immunizing primary infections and vaccines (ε = 0), relatively low vaccination rates are sufficient to achieve zero infections at steady state. However, as immunity becomes more imperfect (larger ε), increasingly high vaccination rates are required to eliminate infections, particularly when the duration of vaccinal immunity is short. This is further emphasized in Fig. 3C, where the minimum vaccination rate ν required to achieve a disease-free state at equilibrium (*37*) is shown as a function of ε for different values of the duration of vaccinal immunity. These results underline that reductions in infection achievable through vaccination are inherently related to the efficacy of the vaccine and the nature of the adaptive immune response (*49*).

We next explore the medium-term dynamic effect of vaccination. Figure 3D shows the ratio of the total number of primary infections during years 1.5 to 5 (i.e., after the vaccine is introduced) relative to the zero vaccination case for different values of the vaccination rate ν and the duration of vaccinal immunity $1/\delta_{vax}$. Figure 3E shows the equivalent for secondary infections. The burden of primary infection decreases with increasing vaccination rate for a given value of vaccinal immunity, $1/\delta_{vax}$. However, for the shortest durations of vaccinal immunity, achievable reductions in the number of secondary cases begin to plateau even for high vaccination rates. This saturation is due to the rapid return of vaccinated individuals to the partially susceptible class if vaccinal immunity is short-lived. Further, if vaccinal immunity wanes very rapidly, vaccination can transiently increase the total number of secondary cases. To further emphasize the dependence of the model results on the vaccination rate and duration of vaccinal immunity, we present time courses of infections and immunity for different durations of vaccinal immunity and vaccination rates in Fig. 3, F to I. In line with intuition, the model illustrates that both high vaccination rates and relatively long durations of vaccine-induced immunity are required to achieve the largest reductions in secondary infection burdens.

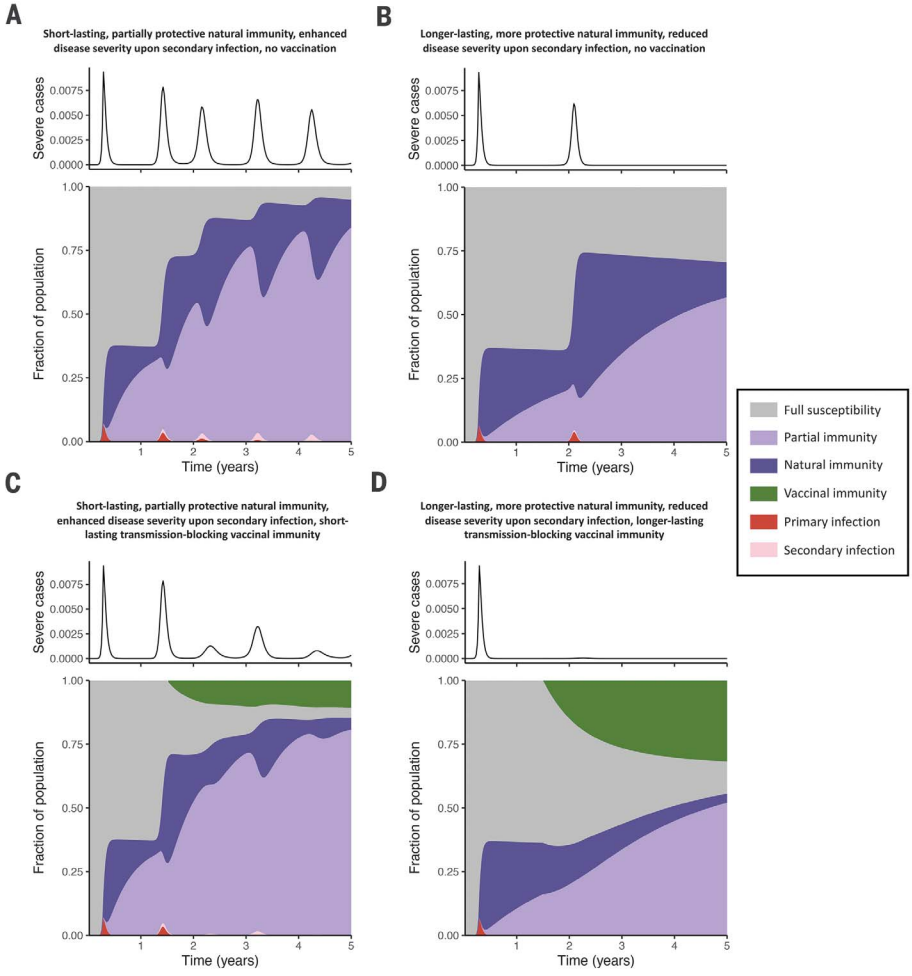

A
Short-lasting, partially protective natural immunity, enhanced disease severity upon secondary infection, no vaccination
B
Longer-lasting, more protective natural immunity, reduced disease severity upon secondary infection, no vaccination
C
Short-lasting, partially protective natural immunity, enhanced disease severity upon secondary infection, short-lasting transmission-blocking vaccinal immunity
D
Longer-lasting, more protective natural immunity, reduced disease severity upon secondary infection, longer-lasting transmission-blocking vaccinal immunity
Severe cases
Fraction of population
Time (years)
Full susceptibility
Partial immunity
Natural immunity
Vaccinal immunity
Primary infection
Secondary infection

Fig. 4. Time series of the fraction of the population with severe primary or secondary cases (top) and area plots of the fraction of the population comprising each immune (S_P, R, S_S, V) or infection (I_P, I_S) class (bottom) over a 5-year time period under four different future scenarios. In all plots, the relative transmissibility of secondary infections (α) is taken to be 1, the fraction of severe primary cases ($x_{sev,p}$) is assumed to be 0.14, a seasonal transmission rate derived from the climate of New York City with no lag is assumed, and a period of social distancing during which the force of infection is reduced to 60% of its original value during weeks 16 to 55 (i.e., the scenario described in Fig. 2B) is enforced. (**A** and **B**) Two scenarios in which no vaccination occurs: a more pessimistic natural immunity scenario, with ε = 0.7, $1/\delta$ = 0.5 years, and 21% of secondary cases being severe (A) and a more optimistic natural immunity scenario, with ε = 0.5, $1/\delta$ = 2 years, and 7% of secondary cases being severe (B). (**C** and **D**) Two scenarios in which vaccination is introduced at a weekly rate ν of 1% at t_{vax} of 1.5 years after the onset of the pandemic: with all the parameters in (A) along with vaccinal immunity lasting $1/\delta_{vax}$ of 0.25 years (C) or with all the same parameters as in (B) along with vaccinal immunity lasting $1/\delta_{vax}$ of 1 year (D).

Infection, disease, and immunity landscape for different possible futures

Figure 4 is a synoptic view of the medium-term impact of vaccination and natural immunity on the immune landscape and incidence of severe disease. We consider four scenarios, assuming seasonal transmission (as in fig. S3) and social distancing according to the pattern depicted in Fig. 2B. Figure 4, A and B, corresponds to futures without vaccination, with Fig. 4A illustrating a more pessimistic scenario of greater susceptibility to secondary infections (ε = 0.7), a relatively short period of natural immunity ($1/\delta$ = 0.5 years), and a greater proportion of severe cases with secondary infection. In contrast, the more optimistic future of Fig. 4B assumes reduced susceptibility to secondary infections (ε = 0.5), a longer duration of natural immunity ($1/\delta$ = 2 years), and a smaller proportion of severe cases with secondary infection. In both cases, the initial pandemic wave is the same, but in the more optimistic scenario (Fig. 4B), natural immunity is longer lasting and, consequently, subsequent infection peaks are delayed. Furthermore, the reduction in susceptibility to secondary infection (smaller ε) in Fig. 4B suppresses the later peaks dominated by secondary infections (Fig. 4A), and substantially less depletion of fully susceptible individuals occurs. In Fig. 4, C and D,

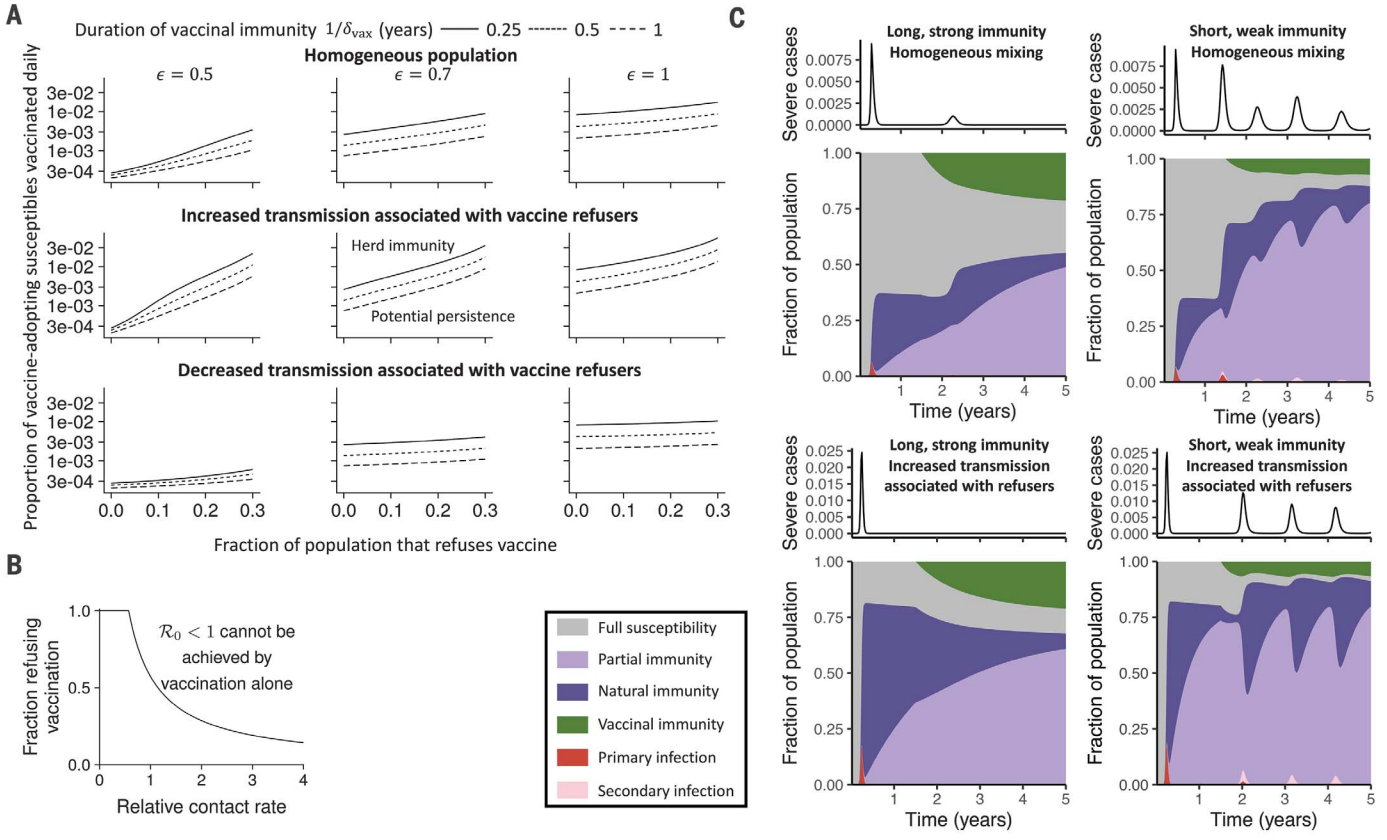


Fig. 5. Effect of vaccine refusal on disease dynamics. (**A**) Daily proportion of vaccine-adopting individuals from the partially and fully susceptible immune classes who must be immunized in order to achieve $\mathcal{R}_0 < 1$ as a function of the fraction of the population that refuses the vaccine (*37*) for different values of the duration of vaccinal immunity ($1/\delta_{vax}$ = 0.25 years, solid line; $1/\delta_{vax}$ = 0.5 years, dashed line; $1/\delta_{vax}$ = 1 year, dotted line) and different values of the susceptibility to secondary infection ϵ [ϵ = 0.5 (left) ϵ = 0.7 (middle) or ϵ = 1 right)]. (Top row) Homogeneous transmission between vaccine adopters and refusers ($c_{11} = c_{12} = c_{21} = c_{22} = 1$). (Middle row) Increased transmission associated with vaccine refusers ($c_{11} = 1$, $c_{12} = 1.25$, $c_{21} = 1.25$, and $c_{22} = 1.5$). (Bottom row) Decreased transmission associated with vaccine refusers ($c_{11} = 1$, $c_{12} = 0.825$, $c_{21} = 0.825$, and $c_{22} = 0.75$). (**B**) Maximum fraction of the population that can refuse vaccination for herd immunity to still be achieved as a function of the contact rate among vaccine refusers c_{22} (*37*). In (A) and (B), the transmission rate is derived from the mean value of seasonal New York City–based weekly reproduction numbers ($\bar{R}_0$ = 1.75) (*37*) (fig. S2C). (**C**) Time series of the fraction of the population with severe primary or secondary cases (top) and area plots of the fraction of the population comprising each immune (S_P, R, S_S, V) or infection (I_P, I_S) class (bottom) over a 5-year time period. The parameters in the left two series are identical to those in Fig. 4C, and the parameters in the right two series are identical to those in Fig. 4D. Additionally, the fraction of the population refusing vaccines is taken to be $N_2 = 0.3$. (Top row) Homogeneous mixing with $c_{11} = c_{12} = c_{21} = c_{22} = 1$. (Bottom row) Increased contacts among vaccine refusers and $c_{11} = 1$, $c_{12} = 1.25$, $c_{21} = 1.25$, and $c_{22} = 1.5$.

these pessimistic and optimistic scenarios are translated into futures with vaccination, which is assumed to be introduced at a weekly rate ν of 1% after a t_{vax} of 1.5 years. The future described in Fig. 4C assumes all the same outcomes as in Fig. 4A and incorporates vaccination with short-lived vaccinal immunity ($1/\delta_{vax}$ = 0.25 years). The future presented in Fig. 4D assumes all the same outcomes as in Fig. 4B in addition to vaccinal immunity lasting for $1/\delta_{vax}$ of 1 year.

Figure 4, C and D, emphasizes the important role that even an imperfect vaccine could have on SARS-CoV-2 dynamics and control [compare with (*7*, *9*, *10*)]. Vaccination substantially reduces subsequent peaks in clinically severe cases, although in the pessimistic future later infection peaks dominated by secondary infections can still occur (Fig. 4C). Furthermore, if a transmission-blocking vaccine confers a relatively long period of protection, and if we make optimistic assumptions regarding the nature of the adaptive immune response (Fig. 4D), a sufficient proportion of fully susceptible individuals can be immunized to suppress future outbreaks within the 5-year time period considered. These trends are qualitatively conserved for different vaccine deployment strategies, such as a pulse of immunization after a t_{vax} of 1.5 years in which a fixed percentage of the fully and partially susceptible populations (S_P and S_S) are vaccinated (fig. S12). However, without sustained immunization strategies, the waning of vaccinal immunity results in a lower susceptible depletion over time and larger future outbreaks relative to the scenarios presented in Fig. 4, C and D.

Impact of heterogeneity

Transmission and clinical heterogeneity

COVID-19 shows marked heterogeneity in transmission and clinical severity with age and other variables (*40*). There are also marked individual heterogeneities, often associated with superspreading events. It is useful to distinguish "environmental" heterogeneity, where high transmission is associated with local environmental (or sociological) factors such as low air exchange, and "intrinsic" heterogeneity, e.g., where certain individuals have consistently higher contact rates (*40*). A number of studies have explored the possibility that intrinsically higher transmission rates for some individuals could reduce the immune threshold for natural or vaccinal herd immunity to COVID-19 (*39*, *50*), echoing classical theory (*51*).

We approximate the impact of intrinsic heterogeneity using a two-subpopulation extension of our homogeneous model (*37*) (figs. S13 to S15). As well as varying transmission between groups, the model assesses covariation between transmission rate and clinical severity. For example, this framing broadly reflects age-structured heterogeneities, in which more clinically threatened older groups might have a lower (because of fewer contacts or possible shielding) or higher [if long-term care facilities are hit (*40*)] transmission rate. We show that moderate heterogeneities do not affect our qualitative projections about the impact of partial natural or vaccinal immunity on epidemic dynamics (figs. S14 and S15). As expected, intrinsic transmission heterogeneity does reduce future burden if there is strong and durable immunity (*39*, *50*) (compare Fig. 4B with figs. S14B and S15B, particularly the subsequent epidemic peaks), because high-transmission individuals would become immune early, reducing average reproduction ratios [and modulating the herd immunity threshold (*39*, *50*)]. However, this impact of intrinsic heterogeneity is weakened [or "buffered" (*35*)] if immunity is imperfect (compare Fig. 4A with figs. S14A and S15A); this is because highly transmissive individuals (e.g., those with a larger social network) contribute proportionately more to secondary transmission when they enter the partially susceptible state. Again, this subtlety illustrates the complexities of even simple variations in immune life history.

Vaccine hesitancy

There is an extensive body of theory on vaccine hesitancy (*52*, *53*). In the homogeneous case, vaccine refusal essentially trades off against vaccine uptake; however, if refusers are spatially or socially clustered, there may be more impact of refusal, both epidemiologically and in terms of the social contagion which underlies it (*52*). We use a simple adaptation of our models (*37*) to explore how transmission heterogeneity influences the epidemiological impact of hesitancy (Fig. 5). As in the vaccination rate titration presented in Fig. 3, B and C, a larger fraction of vaccine refusers in a homogeneous population increases the necessary vaccination rate for herd immunity (Fig. 5A, top). This effect is amplified when the susceptibility to secondary infection is high (compare columns) or the duration of vaccinal immunity is short (compare individual curves). In the heterogeneous case, where vaccine refusers are assumed to have different transmission rates because of higher or lower adherence to NPIs, the minimum vaccination rate to achieve herd immunity is further altered. This can be seen by comparing the rows of Fig. 5A, ordered on the basis of homogeneous contact rates (top), increased contact rates for vaccine refusers (middle), and decreased contact rates for vaccine refusers (bottom). Notably, we find that when vaccine refusers have increased contact rates relative to the rest of the population, vaccination alone may not be able to prevent an outbreak (Fig. 5B). Alternatively, a decrease in contact rates for vaccine refusers decreases their impact. Finally, in Fig. 5C, we reproduce the area plots from Fig. 4, C and D, assuming that 30% of the population refuses the vaccine. This estimate is broadly consistent with recent polls conducted in the United States and Canada (*54*, *55*). We find that the overall disease burden critically depends on the duration and strength of immunity and is larger if vaccine refusers have higher contact rates relative to the rest of the population (compare top and bottom rows).

Caveats

To focus on immune dynamics, we have made several simplifying assumptions. First, we have assumed that transmission of SARS-CoV-2 is seasonal and similar to that of the related β-CoV HCoV-HKU1, although we have also explored the effect of diminished seasonality (*37*). Second, we have simplified the important role for heterogeneities, such as age, clinical severity, transmissibility (*40*), and adaptive immune response (*16*) to primary and secondary (and beyond) infections. Notably, higher viral loads or contact rates in some individuals can lead to superspreading events and heterogeneous transmission patterns (*40*). Additionally, the severity of an infection, especially if associated with higher viremia than in mild cases, could affect the nature of the subsequent adaptive immune response, via antigen-driven expansion of the antibody response (*17*) or exhaustion of the T cell response (*18*). We have explored the effect of these heterogeneities on disease dynamics via a simple model extension (figs. S13 to S15); we find that dynamic impacts of immune variation projected by our homogeneous model are qualitatively robust to these inclusions. Finally, we have considered highly simplified scenarios for NPI adoption and vaccination.

The dynamic impact of these and other parameter variations can be explored interactively at https://grenfelllab.princeton.edu/sarscov2dynamicsplots. For example, strategies to suppress future outbreaks [e.g., (*56*)] could be simulated by increasing the duration and strength of NPIs, then exploring optimal vaccine deployment as vaccines are developed and rolled out. See (*37*) for a full discussion of all caveats and future directions.

Conclusion

We have examined how plausible variations in the natural immune response after SARS-CoV-2 infection and vaccination could interact with seasonal drivers and NPIs to shape the medium-term epidemic dynamics, clinical burden, and immunity landscape to COVID-19. In locations where we expect substantial climatically driven seasonal variation in transmission, such as New York City, the model predicts that a reduction in susceptibility to secondary infection or a longer duration of immunity may lead to a larger secondary infection peak, which may occur earlier if the duration of natural immunity is longer. With smaller annual fluctuations in climate, we find that this nonmonotonic behavior is increasingly suppressed; however, this effect is sensitive to the assumed form of climatic influences on SARS-CoV-2 transmission, which we have taken here to be very similar to those of the related β-CoV HCoV-HKU1. The subsequent pattern of infection peaks is even more sensitive to the relative fraction and transmissibility of primary and secondary cases, as well as the fraction of severe cases for each category. Overall, whereas climatic effects or other seasonal modulators of the transmission rate increase in importance as the pandemic progresses (*25*), our results underline that understanding the immunology of secondary infection (which modulates susceptible supply) is even more dynamically important, especially in the medium term.

The pandemic trajectory can also be substantially altered by mass deployment of vaccines; however, the impact on burden is strongly dependent on the efficacy of the vaccine and the nature of the adaptive immune response. Recent vaccine trials in mice and rhesus macaques indicate the generation of robust immune responses, clinical protection from severe disease, and no evidence of ADE after viral challenge, possibly indicating a more optimistic immune scenario (*44*). Vaccine hesitancy could also decrease vaccination rates (*53*), leading to lower levels of population immunity. Nevertheless, even with imperfect vaccinal immunity and moderate vaccination rates, our results indicate that vaccination may accelerate pandemic control. Ultimately, quantitatively projecting the impact of vaccination, antivirals, and therapeutics will require more granular immuno-epidemiological models; however, parameterizing such models will continue to present huge challenges for this novel virus. We argue that a family of simple and more complex models, with a careful focus on model comparison and averaging, is the way ahead (*57*).

Our work underlines that relying on the status of infection of an individual as the main observable during an ongoing epidemic is insufficient to characterize the complex immune landscape generated by the pandemic. This is in line with ongoing calls for the development of a Global Immunological Observatory for the surveillance of population-level susceptibility and immunity to circulating pathogens, as well as the emergence of new strains (*58*–*60*). Given the increasingly recognized importance of both

T cell–mediated (*14*, *15*) and antibody-mediated (*11*–*13*) adaptive immune responses in the recovery from SARS-CoV-2 infection, regular testing of antibody presence, and correlates of protection such as neutralization, as well as T cell immunity, in parallel with viral testing, will be required to adequately characterize population-level natural and vaccinal immunity to this pathogen. Specifically, our model indicates a key need to establish (i) the duration and strength of transmission-blocking and clinical immunity after primary (and subsequent) infection and vaccination; (ii) population and individual variations in these parameters (age, sex, etc.); and (iii) the impact of viral evolution, coinfection, and other pathogen characteristics on COVID-19 infection and disease. Quantifying these parameters will require long-term major investments in integrated viral and immune surveillance. Moving beyond the current pandemic, these structures (and associated developments in biology, informatics, and translation) will be powerful bases for understanding and combating inevitable future microbial threats (*58*–*60*).

This work emphasizes the complex dependence of the immune landscape generated by SARS-CoV-2 infection on the presently uncertain nature of the adaptive immune response to this virus and the efficacy of potential future vaccines. Depending on how these unfold, the model predictions for future clinical burdens range from sustained epidemics to near–case elimination. Consequently, accurately characterizing the individual immune life histories and the cumulative immune landscape of the population to SARS-CoV-2 primary and secondary infection and vaccination will be critical for the management and control of the ongoing pandemic.

REFERENCES AND NOTES

1. B. M. Laksono, R. D. de Vries, S. McQuaid, W. P. Duprex, R. L. de Swart, *Viruses* **8**, 210 (2016).
2. P. D. Ellner, *Infection* **26**, 263–269 (1998).
3. J. Dushoff, J. B. Plotkin, S. A. Levin, D. J. D. Earn, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16915–16916 (2004).
4. V. E. Pitzer *et al.*, *PLOS Pathog.* **11**, e1004591 (2015).
5. N. Ferguson, R. Anderson, S. Gupta, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 790–794 (1999).
6. B. S. Graham, *Science* **368**, 945–946 (2020).
7. S. Gandon, M. J. Mackinnon, S. Nee, A. F. Read, *Nature* **414**, 751–756 (2001).
8. G. Anichini *et al.*, *Vaccines* **8**, 66 (2020).
9. N. Arinaminpathy *et al.*, *Am. J. Epidemiol.* **186**, 92–100 (2017).
10. T. A. Perkins *et al.*, *PLOS Comput. Biol.* **15**, e1006710 (2019).
11. J. Huang, T. Mao, S. Li, L. Wu, X. Xu, H. Li, C. Xu, F. Su, J. Dai, J. Shi, J. Cai, C. Huang, X. Lin, D. Chen, X. Lin, B. Sun, S. Tang, Long period dynamics of viral load and antibodies for SARS-CoV-2 infection: An observational cohort study. medRxiv 20071258 [Preprint]. 27 April 2020. https://doi.org/10.1101/2020.04.22.20071258.
12. N. M. A. Okba *et al.*, *Emerg. Infect. Dis.* **26**, 1478–1488 (2020).
13. W. Tan, Y. Lu, J. Zhang, J. Wang, Y. Dan, Z. Tan, X, He, C. Qian, Q. Sun, Q. Hu, H. Liu, S. Ye, X. Xiang, Y. Zhou, W. Zhang, Y. Guo, X.-H. Wang, W. He, X. Wan, F. Sun, Q. Wei, C. Chen, G. Pan, J. Xia, Q. Mao, Y. Chen, Viral Kinetics and Antibody Responses in Patients with COVID-19. medRxiv 20042382 [Preprint]. 26 March 2020. https://doi.org/10.1101/2020.03.24.20042382.
14. A. Grifoni *et al.*, *Cell* **181**, 1489–1501.e15 (2020).
15. J. Braun, L. Loyal, M. Frentsch, D. Wendisch, P. Georg, F. Kurth, S. Hippenstiel, M. Dingledey, B. Kruse, F. Fauchere, E. Baysal, M. Mangold, L. Henze, R. Lauster, M. Mall, K. Beyer, J. Roehmel, J. Schmitz, S. Miltenyi, M. A. Mueller, M. Witzenrath, N. Suttorp, F. Kern, U. Reimer, H. Wenschuh, C. Drosten, V. M. Corman, C. Gisecke-Theil, L.-E. Sander, A. Thiel, Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. medRxiv 20061440 [Preprint]. 22 April 2020. https://doi.org/10.1101/2020.04.17.20061440.
16. D. Mathew *et al.*, *Science* **369**, eabc8511 (2020).
17. Y. Wang *et al.*, *J. Clin. Invest.* 10.1172/JCI138759 (2020).
18. S. De Biasi *et al.*, *Nat. Commun.* **11**, 3434 (2020).
19. Q.-X. Long *et al.*, *Nat. Med.* **26**, 845–848 (2020).
20. D. F. Gudbjartsson *et al.*, *N. Engl. J. Med.* 10.1056/NEJMoa2026116 (2020).
21. S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, M. Lipsitch, *Science* **368**, 860–868 (2020).
22. M. R. Macnaughton, *Infect. Immun.* **38**, 419–423 (1982).
23. K. A. Callow, H. F. Parry, M. Sergeant, D. A. J. Tyrrell, *Epidemiol. Infect.* **105**, 435–446 (1990).
24. E. R. Gaunt, A. Hardie, E. C. Claas, P. Simmonds, K. E. Templeton, *J. Clin. Microbiol.* **48**, 2940–2947 (2010).
25. R. E. Baker, W. Yang, G. A. Vecchi, C. J. E. Metcalf, B. T. Grenfell, *Science* **369**, 315–319 (2020).
26. W. Liu *et al.*, *J. Infect. Dis.* **193**, 792–795 (2006).
27. W.-C. Cao, W. Liu, P.-H. Zhang, F. Zhang, J. H. Richardus, *N. Engl. J. Med.* **357**, 1162–1163 (2007).
28. O.-W. Ng *et al.*, *Vaccine* **34**, 2008–2014 (2016).
29. K. W. Ng *et al.*, Pre-existing and de novo humoral immunity to SARS-CoV-2 in humans. bioRxiv 095414 [Preprint]. 23 July 2020. https://doi.org/10.1101/2020.05.14.095414.
30. J. A. Tetro, *Microbes Infect.* **22**, 72–73 (2020).
31. M. Jaume *et al.*, *Hong Kong Med. J.* **18** (suppl. 2), 31–36 (2012).
32. M. S. Yip *et al.*, *Virol. J.* **11**, 82 (2014).
33. M. S. Yip *et al.*, *Hong Kong Med. J.* **22** (suppl. 4), 25–31 (2016).
34. A. D. Becker, B. T. Grenfell, *PLOS ONE* **12**, e0185528 (2017).
35. S. E. Morris *et al.*, *J. R. Soc. Interface* **12**, 20141245 (2015).
36. V. E. Pitzer *et al.*, *Science* **325**, 290–294 (2009).
37. Materials and methods are available as supplementary materials.
38. J. Bedford *et al.*, *Lancet* 10.2139/ssrn.3678593 [Preprint] (2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3678593
39. T. Britton, F. Ball, P. Trapman, *Science* **369**, 846–849 (2020).
40. B. M. Althouse, E. A. Wenger, J. C. Miller, S. V. Scarpino, A. Allard, L. Hebert-Dufresne, H. Hu, Stochasticity and heterogeneity in the transmission dynamics of SARS-CoV-2. arXiv:2005.13689 [q-bio.PE] (27 May 2020).
41. S. W. Park *et al.*, *J. R. Soc. Interface* **17**, 20200144 (2020).
42. Z. Wu, J. M. McGoogan, *JAMA* **323**, 1239–1242 (2020).
43. V. O. Puntmann *et al.*, *JAMA Cardiol.* (2020).
44. M. J. Mulligan, K. E. Lyke, N. Kitchin, J. Absalon, A. Gurtman, S. P. Lockhart, K. Neuzil, V. Raabe, R. Bailey, K. A. Swanson, P. Li, K. Koury, W. Kalina, D. Cooper, C. Fonter-Garfias, P.-Y. Shi, O. Tuereci, K. R. Tompkins, E. E. Walsh, R. Frenck, A. R. Falsey, P. R. Dormitzer, W. C. Gruber, U. Sahin, K. U. Jansen, Phase 1/2 Study to Describe the Safety and Immunogenicity of a COVID-19 RNA Vaccine Candidate (BNT162b1) in Adults 18 to 55 Years of Age: Interim Report. medRxiv 20142570 [Preprint]. 1 July 2020. https://doi.org/10.1101/2020.06.30.20142570.
45. N. van Doremalen *et al.*, *Nature* 10.1038/s41586-020-2608-y (2020).
46. U.S. Centers for Disease Control and Prevention, "Final estimates for 2009–10 Seasonal Influenza and Influenza A (H1N1) 2009 Monovalent Vaccination Coverage—United States, August 2009 through May, 2010" (2011); https://www.cdc.gov/flu/fluvaxview/coverage_0910estimates.htm.
47. H. Gilmour, N. Hofmann, H1N1 Vaccination (Statistics Canada, 2015); https://www150.statcan.gc.ca/n1/pub/82-003-x/2010004/article/11348-eng.htm.
48. Assuming a simplified scenario in which a constant proportion of the total population is vaccinated weekly, the proportion of the population that is not vaccinated, U, can be modeled to decrease exponentially in time as $U(t) = \exp(-\nu t)$. Consequently, in order to achieve 27% vaccination coverage after an 8-month (or 34.7-week) period [i.e., U (34.7) = 0.73], a weekly vaccination rate of 0.91% of the remaining unvaccinated population would be required.
49. M. E. Halloran *et al.*, *Design and Analysis of Vaccine Studies* (Springer, 2010).
50. R. Aguas, R. M. Corder, J. G. King, G. Goncalves, M. U. Ferreira, M. G. M. Gomes, Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics. medRxiv 20160762 [Preprint]. 31 August 2020. https://doi.org/10.1101/2020.07.23.20160762.
51. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, 1991).
52. M. Salathé, S. Bonhoeffer, *J. R. Soc. Interface* **5**, 1505–1508 (2008).
53. E. Dubé *et al.*, *Hum. Vaccin. Immunother.* **9**, 1763–1773 (2013).
54. S. M. O'Keefe, "One in Three Americans Would Not Get COVID-19 Vaccine" (2020; https://news.gallup.com/poll/317018/one-three-americans-not-covid-vaccine.aspx).
55. K. Frank, R. Arim, "Canadians' willingness to get a COVID-19 vaccine when one becomes available: What role does trust play?"(2020; https://www150.statcan.gc.ca/n1/pub/45-28-0001/2020001/article/00043-eng.htm).
56. M. G. Baker, A. Kvalsvig, A. J. Verrall, *Med. J. Australia* (2020); https://www.mja.com.au/journal/2020/new-zealands-covid-19-elimination-strategy.
57. C. Viboud *et al.*, *Epidemics* **22**, 13–21 (2018).
58. C. J. E. Metcalf *et al.*, *Lancet* **388**, 728–730 (2016).
59. C. J. E. Metcalf, M. J. Mina, A. K. Winter, B. T. Grenfell, *Lancet* **389**, 252 (2017).
60. M. J. Mina *et al.*, *eLife* **9**, e58989 (2020).
61. C. M. Saad-Roy *et al.*, Code for immune life-history, vaccination, and the dynamics of SARS-CoV-2 over the next five years. Zenodo (2020; doi: 10.5281/zenodo.4025265).
62. Created with BioRender.com.

ACKNOWLEDGMENTS

We acknowledge useful discussions with B.Thompson and the Wellcome COVID-19 Futures Group and the members of the DELVE Immunology group. **Funding:** C.M.S.-R. acknowledges support from the Natural Sciences and Engineering Research Council of Canada through a Postgraduate-Doctoral Scholarship. C.E.W. is an Open Philanthropy Project fellow of the Life Sciences Research Foundation. R.E.B. is supported by the Cooperative Institute for Modelling the Earth System (CIMES). S.A.L. and C.M.S.-R. acknowledge support from the James S. McDonnell Foundation 21st Century Science Initiative Collaborative Award in Understanding Dynamic and Multi-scale Systems. S.A.L. acknowledges support from the C3.ai Digital Transformation Institute, the National Science Foundation under grant CNS-2027908, and the National Science Foundation Expeditions Grant CCF1917819. B.T.G. acknowledges support from the U.S. CDC and Flu Lab. **Author contributions:** C.M.S.-R., C.E.W., C.J.E.M., and B.T.G. designed the study. C.M.S.-R. and C.E.W. performed the simulations and wrote the manuscript. C.M.S.-R., C.E.W., R.E.B., C.J.E.M., and B.T.G. analyzed the results. S.E.M. developed the Shiny application. C.M.S.-R., C.E.W., and S.A.L. developed the analytical equilibrium analysis. All authors contributed to interpreting the results and editing the manuscript. **Competing interests:** The authors have no competing interests. **Data and materials availability:** This manuscript contains no new data, and all referenced data sets are publicly available. Code for all analysis is available at Zenodo (*61*). This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/811/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S15
References (*63*–*70*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

8 July 2020; accepted 16 September 2020
Published online 21 September 2020
10.1126/science.abd7343

PLANT SCIENCE

Cell wall remodeling and vesicle trafficking mediate the root clock in *Arabidopsis*

Guy Wachsman[1,2], Jingyuan Zhang[2], Miguel A. Moreno-Risueno[3], Charles T. Anderson[4], Philip N. Benfey[1,2]*

In *Arabidopsis thaliana*, lateral roots initiate in a process preceded by periodic gene expression known as the root clock. We identified the vesicle-trafficking regulator GNOM and its suppressor, ADENOSINE PHOSPHATE RIBOSYLATION FACTOR GTPase ACTIVATION PROTEIN DOMAIN3, as root clock regulators. GNOM is required for the proper distribution of pectin, a mediator of intercellular adhesion, whereas the pectin esterification state is essential for a functional root clock. In sites of lateral root primordia emergence, both esterified and de-esterified pectin variants are differentially distributed. Using a reverse-genetics approach, we show that genes controlling pectin esterification regulate the root clock and lateral root initiation. These results indicate that the balance between esterified and de-esterified pectin states is essential for proper root clock function and the subsequent initiation of lateral root primordia.

Both plants and animals use biological clocks to estimate time and space during development. In vertebrates, the formation of presumptive vertebrae, known as somitogenesis, is one of the best characterized mechanisms involving periodic gene expression. In plants, before the initiation of lateral root primordia (LRP), the auxin response marker DR5 and other genes exhibit periodic expression in a region called the oscillation zone (*1*, *2*). This rhythmic gene expression, known as the root clock, produces prebranch sites, DR5-expressing spots that develop into LRP (*1*). The root clock represents a flexible oscillator, producing prebranch sites approximately every 6 hours. Most genes implicated in the initiation of LRP are related to auxin signaling (*3*, *4*) or metabolism (*5*). Recently, it has been shown that auxin-dependent changes in endodermis cells are required before the initiation of LRP in underlying pericycle cells (*6*). One candidate protein that putatively mediates signaling between cells and is also involved in LR initiation is GNOM (*7*).

GNOM is a trans-Golgi network-localized (*8*) ARF-GEF [ADP RIBOSYLATION FACTOR GUANINE NUCLEOTIDE EXCHANGE FACTOR (*9*)] that regulates vesicle trafficking to the plasma membrane. It had been suggested that the reduced LR phenotype of the weak *gnom* allele, *fwr*, is mediated by one of the PIN-FORMED (PIN) proteins. However, the *fwr* mutant has properly localized PIN1 (*7*). In addition, there are no known PIN mutants with LR phenotypes. These data suggest that the lack of LRs in the *gnom* mutant might depend on other factors that are not directly involved in auxin signaling. Vesicle trafficking is necessary for the secretion of cell wall components including pectin, hemicelluloses, and cell wall–modifying enzymes (*10*). Here, we show that disrupting GNOM alters the distribution of pectin and blocks the functioning of the root clock. Furthermore, higher levels of esterified and de-esterified pectin are differentially localized around emerging LRP, whereas modulation of either form of pectin leads to a loss of prebranch site formation. On the basis of these data, we conclude that GNOM-dependent vesicle trafficking mediates the distribution of pectin in the cell wall, which in turn is essential for root clock function and subsequent LR formation.

[1]Howard Hughes Medical Institute, Duke University, Durham, NC 27708, USA. [2]Department of Biology, Duke University, Durham, NC 27708, USA. [3]Centro de Biotecnología y Genómica de Plantas (Universidad Politécnica de Madrid – Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria), 28223 Pozuelo de Alarcón (Madrid), Spain. [4]Department of Biology, The Pennsylvania State University, University Park, PA, 16802, USA.
*Corresponding author. Email: philip.benfey@duke.edu

LRP cell divisions initiate soon after oscillation

The root clock is defined as a 6-hour periodic DR5 expression in a region known as the oscillation zone near the border of the elongation and differentiation zones of the *Arabidopsis thaliana* root (*1*, *2*). After peak DR5-driven luciferase expression, the region shrinks to form a confined spot defined as a prebranch site, where LRP emerge subsequently. All LRP emerge from prebranch sites, and most prebranch sites develop into LRP (*1*). To better understand the relationship between LR initiation and prebranch site formation, we examined roots expressing both the DR5::LUC reporter, which marks prebranch sites, and pCLE44::GFP, which marks early LRP (fig. S1, A and B; see below for details on the CLE44 marker). In most cases (21/26), the most recent DR5::LUC-marked prebranch site was associated with an early-stage LRP. We conclude that LRP initiation starts concurrent with or soon after the formation of a prebranch site.

RNA-sequencing screen identifies oscillating genes

To identify genes with elevated expression in the oscillation zone and/or in the most recently formed prebranch site, we used RNA-sequencing (RNA-seq) analysis of sections from individual plants using the modified Smart-seq2 method (*11*). In a spatial experiment, we performed RNA-seq on transverse sections from the oscillation zone, the first prebranch site, and their flanking regions (Fig. 1A). In a complementary, temporal experiment, we performed RNA-seq on DR5::LUC-expressing sections before, during, and after peak expression (Fig. 1B and supplementary materials). To validate the quality of the data, we compared RNA-seq reads of the luciferase transgene (driven by the DR5 promoter) with luciferase imaging and found a high correlation (Fig. 1, C and D). Multidimensional scaling analysis and other statistical analyses provided further evidence for the fidelity of the data (Fig. 1E and fig. S2). To identify putative root clock regulators, we selected genes that were overrepresented in the oscillation zone-spatial, the oscillation zone-spatial together with the oscillation zone-temporal, or the prebranch site-spatial datasets (Fig. 1, F to I, and table S1, A to D) and used these three lists to perform gene ontology (GO) analyses (table S2). We identified CLE44 as one of the genes expressed in incipient prebranch sites with similar frequency as DR5 (fig. S1, C to E, and movie S1) and used it as a prebranch site marker together with DR5 in subsequent experiments. Some of the most common GO terms were related to cell wall biogenesis and vesicle trafficking (Fig. 2A). Of particular interest was a term for pectin-metabolic process, including genes that code for pectin lyases, pectate lyases, pectin methylesterases (PMEs), and their inhibitors (PMEIs) ($P = 0.005$; Fig. 2, B and C, and table S2). PMEs de-methyl-esterify galacturonic acid residues of homogalacturonan pectin. Pectin is synthesized in the Golgi in a highly methyl-esterified form and delivered to the apoplast, where it can be de-methyl-esterified. Blockwise de-esterified homogalacturonan chains can be crossed-linked by calcium, leading to stiffening of the cell wall and/or promoting adhesion between cells, whereas randomly de-esterified homogalacturonan can be degraded by pectinases, resulting in wall loosening and/or cell separation (*12*, *13*). All nine oscillation zone-spatial PMEs belong to group 1 (*14*) within the type I class of PMEs, which contain a pro domain at the N terminus that requires preprocessing in the Golgi before secretion to the apoplast (*15*). Another group of genes overrepresented in the oscillation zone-spatial dataset is involved in auxin transport and

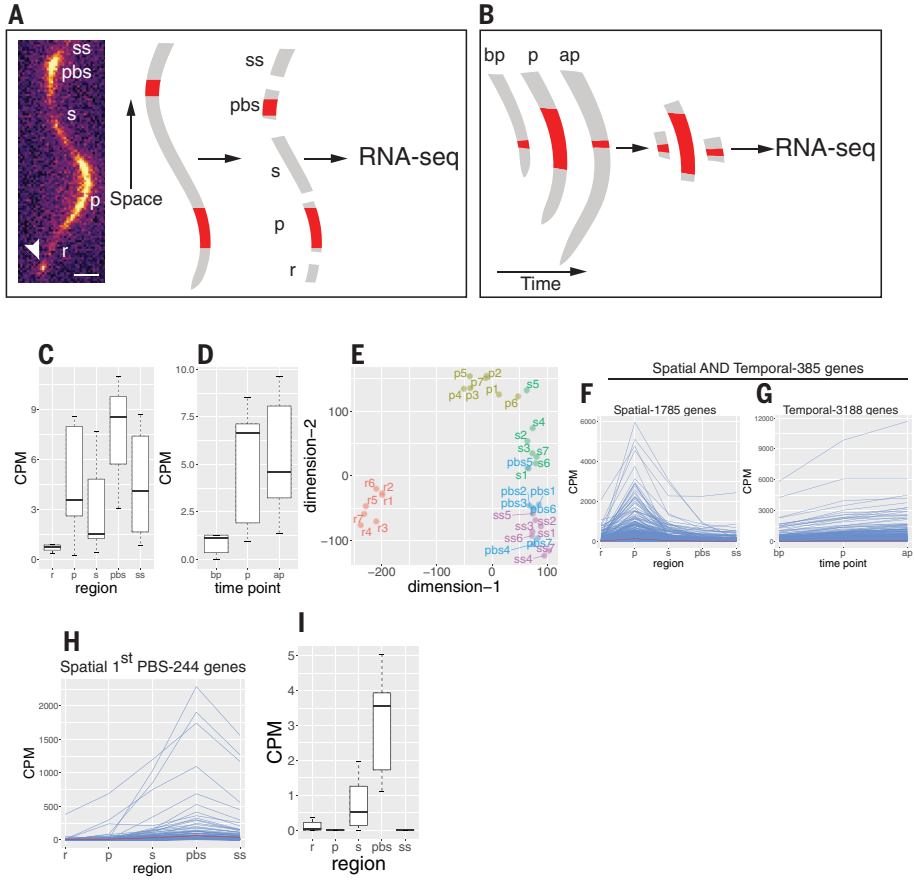


Fig. 1. Experimental design and analysis of RNA-seq experiments. (**A**) Spatial experiment. Seven DR5::LUC roots were imaged using the Lumazone system (left). Once peak expression was detected in the oscillation zone, the root was dissected into five sections: r, a section proximal to the root meristem and distal to the oscillation zone; p, the oscillation zone; s, a section proximal to the oscillation zone and distal to the youngest prebranch site; pbs, the youngest prebranch site; and ss, a section proximal to the prebranch site. Each section was used to generate a transcriptome profile (35 total). Arrowhead indicates the root tip. Scale bar, 0.3 mm. (**B**) Same as in (A) but a root section was collected from the oscillation zone before peak expression (bp), during peak expression (p), or after peak expression (ap). (**C** and **D**) CPM reads for luciferase RNA in the spatial experiment (C) and the temporal experiment (D). (**E**) Multidimensional scaling plot of the spatial experiment. Note that the regions (r, p, s, pbs, and ss) mostly form distinctive clades. (**F**) Expression pattern of all 1785 genes in the spatial experiment with a $\log_2$FC > 0.5 (FC > 1.414) in the p region compared with each of the other zones. (**G**) Expression pattern of all 3188 genes in the temporal experiment with a $\log_2$FC > 0.5 (FC > 1.414) in the p region (peak luciferase expression) compared with the bp region (before peak luciferase expression). (**H**) Expression pattern of all 244 genes in the spatial experiment with a $\log_2$FC > 0.5 (FC > 1.414) in the pbs zone compared with each of the other zones. (**I**) Expression profile of CLE44, which was later used as a prebranch site marker. CPM, counts per million.

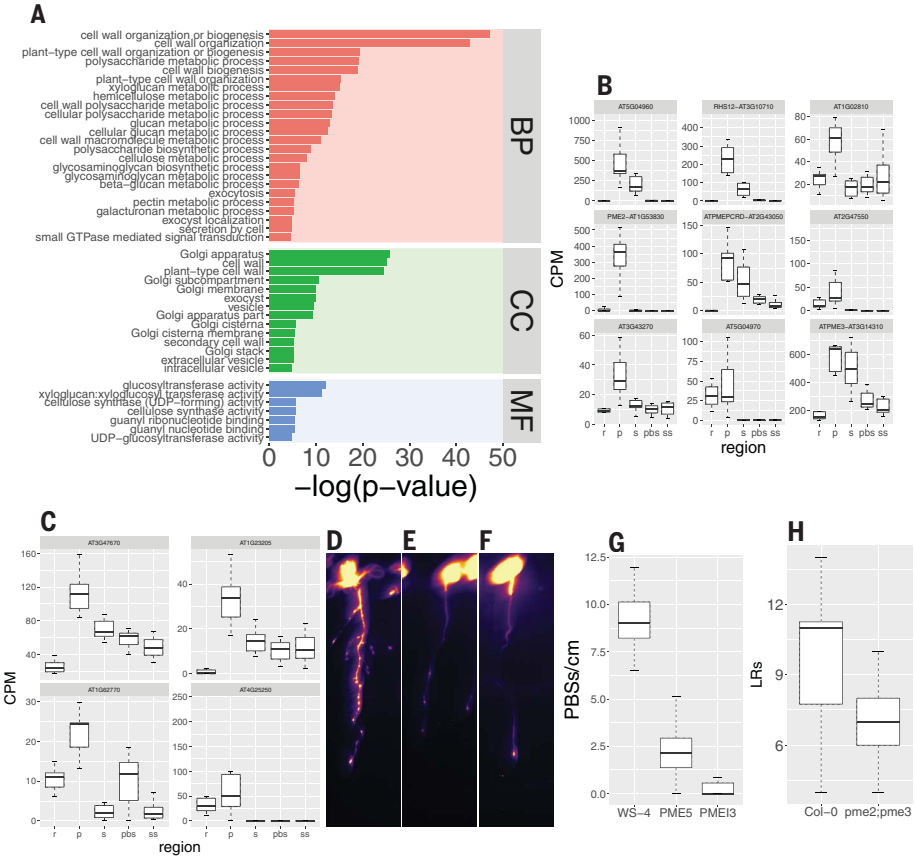


Fig. 2. Cell wall-, vesicle trafficking-, and membrane-related genes are overrepresented in the oscillation zone and prebranch site and required for its formation. (**A**) GO term analysis of the 1785 genes described in Fig. 1F (BP, biological process; CC, cellular component; MF, molecular function). The plot depicts vesicle trafficking-, Golgi-, cell wall-, small GTPase–mediated signal transduction-, and carbohydrate metabolism-related GO terms with $P \leq 0.01$ (hypergeometric test using the gProfiler R package; see table S2 for the complete list). (**B** and **C**) Expression pattern of nine PME (B) and four PMEI (C) genes in the spatial experiment. (**D** to **F**) pCLE44::LUC expression in WT [WS-4, (D)], PME5-overexpressing (E), and PMEI3-overexpressing (F) seedlings after EtOH induction. Scale bar, 0.2 cm. (**G**) Quantification of prebranch site number [(D) to (F)]. $P < 10^{-14}$, Wilcoxon rank sum test. (**H**) LR number of *pme2;pme3* compared with Col-0 WT. $P = 0.007$, Wilcoxon rank sum test.

signal transduction. Mutations in five of these genes (LAX3, AUX1, ARF19, IAA14, and IAA28) directly affect LR formation [fig. S3A and (*3*, *16*–*18*)]. A third group overrepresented in the oscillation zone-spatial dataset is involved in the transport, metabolism, and response to nitrogen-related compounds (fig. S3B), all processes that have been shown to affect LR density and outgrowth (*19*, *20*). For example, the tonoplast aquaporin *TONOPLAST INTEGRAL PROTEIN2;1/DELTA-TIP1* (*TIP2;1*), which is involved in NH_3 transport to the vacuole and response to NH_3 accumulation, is redundantly required, together with two other TIP proteins, for LR initiation (*21*, *22*). We also identified two additional genes that were not included in the dataset because they did not fulfill the log-fold change (logFC) > 0.5 criteria but still had high expression in the oscillation zone: AMT2;1, an ammonium transporter required for ammonium-dependent LR branching (*20*), and TGA4, a basic leucine zipper transcription factor controlling nitrate-dependent LR initiation (*19*). Another group of genes with high expression in the oscillation zone belong to the cysteine-rich receptor-like protein kinase (crRLK) subfamily (fig. S3C). FERONIA, one of the members of this family, was recently shown to bind pectin and to regulate pavement cell morphogenesis (*23*). Finally, our approach also identified 10 genes in a GO term for LR morphogenesis (fig. S3D), which involve auxin transport (LAX3 and AUX1), auxin signaling (ARF11/ARF19, IAA14 and IAA28), exocytosis (EXO70A1), and the promotion of protein degradation by ubiquitin ligase activity (XBAT32). Together, these results suggest that the oscillation zone functions as a transcriptional hub for multiple pathways required for emergence of LRP.

PMEs and PMEIs are required for prebranch site and LR formation

Because PME5 and PMEI3 have been implicated in phyllotaxis (*24*), we decided to explore the role of pectin esterification in LR branching. To determine whether changing the pectin esterification state also affects rhyzotaxis, we inducibly overexpressed PME5 and PMEI3 and found a severe reduction in prebranch site number in both cases (Fig. 2, D to G, and fig. S4, A to D), as well as a decrease in LR number (fig. S4, E to G). On the basis of our RNA-seq results, we identified two PME genes, the mutants of which, *pme2* and *pme3*, as well as the double mutant combination, showed reduced numbers of LRs (Fig. 2H and fig. S4, H to K). These results suggest that the balance between pectin esterification and de-esterification underpins the functionality of the root clock.

Vesicle-trafficking genes mediate root clock and LR formation

To complement our reverse-genetics approach, we mutagenized DR5::LUC seeds and screened

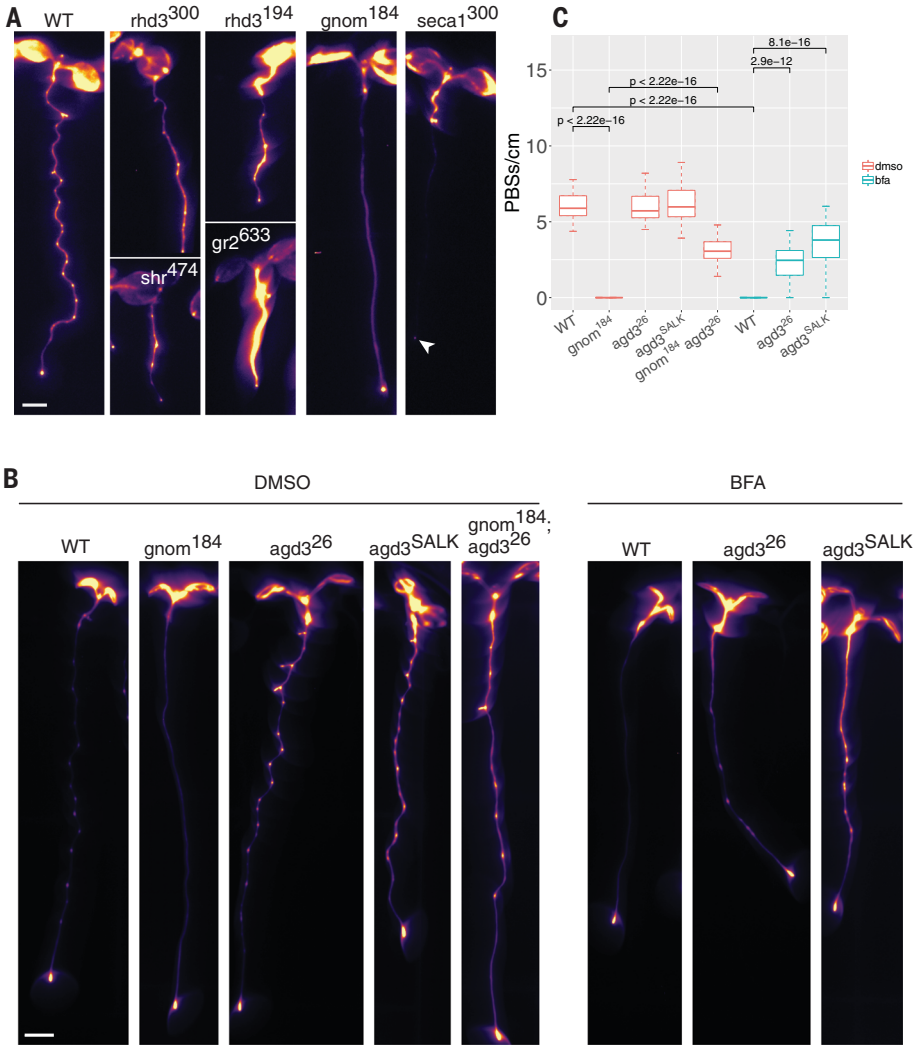


Fig. 3. Perturbation of vesicle trafficking affects the root clock. (**A**) Root clock phenotype of six ethyl methanesulfonate mutants. Line name is shown as a superscript and the gene name is in lowercase letters. Arrowhead indicates DR5::LUC expression in the root tip. (**B**) Expression of pCLE44::LUC in different genotypes under mock treatment [dimethyl sulfoxide (DMSO), left five panels] or BFA treatment (right three panels). (**C**) Quantification of pCLE44::LUC prebranch site number in *gnom*^{184} and its suppressor, *agd3* (SALK and the EMS mutant *agd3*^{26}), with or without 3 μM BFA (*t* test). Scale bar, 0.2 cm.

for alterations in root clock function based on imaging. We identified six mutations with modified root clock function (table S3A) and mapped the affected genes [table S3 and (*25*)]. Two independent mutations were found in *ROOT HAIR DEFECTIVE 3 (RHD3)* and one each in *GNOM*, *SHORTROOT*, *GLUTHATIONE REDUCTASE*, and *SECA1* (Fig. 3A and table S3). Because half of the mutations affect genes coding for vesicle-trafficking-related proteins (RHD3 and GNOM) and the *gnom* mutant allele (*gnom*^{184}) did not appear to be highly pleiotropic, we focused our analysis on this mutant (Fig. 3 and fig. S5, A to I). This allele could be partially complemented by its native promoter and 3′ untranslated region or with the G1090-inducible overexpression promoter [fig. S5J and (*26*)]. To better understand the *gnom*^{184} phenotype, we grew wild-type (WT) seedlings on brefeldin A (BFA), a vesicle-trafficking inhibitor that targets a subset of plant ARF-GEFs, including GNOM (*27*). A single amino acid substitution in the catalytic Sec7 domain of GNOM can convert WT seedlings to BFA-resistant seedlings (*27*), which produce LRs despite exposure to BFA, suggesting that BFA specifically targets GNOM. Seedlings grown on low BFA concentrations (3 to 5 μM) have a similar phenotype to *gnom*^{184} with no LRs [fig. S5I and (*27*)]. We next tested the effect of BFA on the root clock using the DR5::LUC and the pCLE44::LUC lines. The expression pattern of these markers was strongly altered, phenocopying the *gnom*^{184} allele (Fig. 3, B and

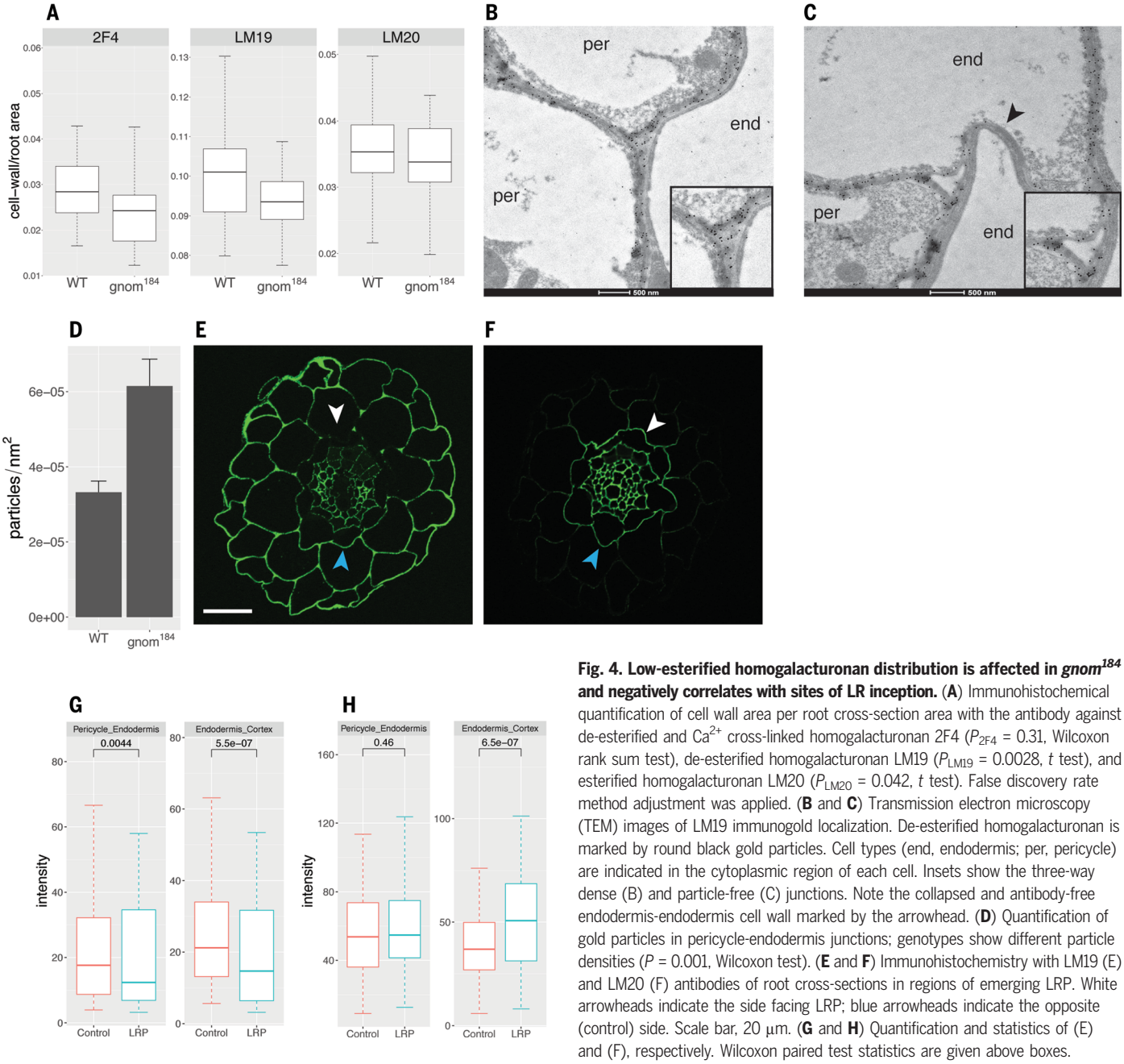


Fig. 4. Low-esterified homogalacturonan distribution is affected in *gnom*^{184} and negatively correlates with sites of LR inception. (**A**) Immunohistochemical quantification of cell wall area per root cross-section area with the antibody against de-esterified and Ca^{2+} cross-linked homogalacturonan 2F4 (P_{2F4} = 0.31, Wilcoxon rank sum test), de-esterified homogalacturonan LM19 (P_{LM19} = 0.0028, *t* test), and esterified homogalacturonan LM20 (P_{LM20} = 0.042, *t* test). False discovery rate method adjustment was applied. (**B** and **C**) Transmission electron microscopy (TEM) images of LM19 immunogold localization. De-esterified homogalacturonan is marked by round black gold particles. Cell types (end, endodermis; per, pericycle) are indicated in the cytoplasmic region of each cell. Insets show the three-way dense (B) and particle-free (C) junctions. Note the collapsed and antibody-free endodermis-endodermis cell wall marked by the arrowhead. (**D**) Quantification of gold particles in pericycle-endodermis junctions; genotypes show different particle densities (P = 0.001, Wilcoxon test). (**E** and **F**) Immunohistochemistry with LM19 (E) and LM20 (F) antibodies of root cross-sections in regions of emerging LRP. White arrowheads indicate the side facing LRP; blue arrowheads indicate the opposite (control) side. Scale bar, 20 μm. (**G** and **H**) Quantification and statistics of (E) and (F), respectively. Wilcoxon paired test statistics are given above boxes.

C, and fig. S5, G and H). To further identify root-clock-regulated genes that function in the GNOM pathway, we used the mutagenized DR5::LUC population and screened for seedlings that could develop LRs under BFA treatment. We identified two BFA-insensitive mutants (table S3) displaying both LRs and prebranch sites upon BFA treatment. The first affected gene, *LEUCINE CARBOXYLMETHYLTRANSFERASE1/SUPPRESSOR OF BRI1* (*LCMT/SBI1*), is involved in terminating the brassinosteroid signaling cascade during endocytosis by promoting degradation of the hormone receptor BRI1 [fig. S5, K to M, and (*28*)]. The second gene, *ARF-GTP ACTIVATING PROTEIN DOMAIN* (*AGD3*), is required for scission during vesicle formation (*29*). GNOM is an ARF-GEF that facilitates the exchange of GDP to GTP, leading to membrane recruitment of the ARF protein. AGD3 acts in the opposite direction by activating ARF-dependent GTP hydrolysis, allowing vesicle budding and eventually membrane dissociation of ARF-GDP (*30*). When introduced into the *gnom*^{184} background, our mutant *agd3* allele (*agd3*^{26}) partially rescued the loss of root clock function and the LR phenotype (Fig. 3, B and C, and fig. S5, G to I). These results indicate that the vesicle-trafficking GNOM-AGD3 circuit is required for LRP initiation and root clock function.

GNOM regulates homogalacturonan distribution

Because pectic homogalacturonan is synthesized in its esterified form in the Golgi (covalently linked to a methyl group through an ester bond) and then secreted to the cell wall, where it can be de-esterified by PMEs, we sought to determine whether its involvement in the root clock and LRP initiation requires GNOM-dependent vesicle trafficking. We used immunohistochemistry to determine whether

gnom^{184} has altered distribution of three homogalacturonan forms (*31*): de-esterified and Ca^{2+} cross-linked [2F4 antibody; (*32*)], de-esterified (LM19 antibody), and esterified [LM20 antibody; (*33*)]. The mean intensity signal of homogalacturonan in all three experiments did not differ between WT and *gnom*^{184} (fig. S6). Using xyloglucanase to digest hemicellulose did not unmask additional pectin sites or change the intensity of the antipectin antibody staining (fig. S6), as previously shown for mannan polysaccharide (*34*). However, the cell wall area occupied by the low-esterified form was significantly lower in *gnom*^{184} than in WT (P = 0.0028; Fig. 4A). A severe *gnom* allele has been shown to alter the distribution of de-esterified homogalacturonan in the cell wall (*35*). Because the localization of de-esterified pectin is altered in *gnom*^{184} and the boundary between the pericycle and the endodermis is relevant for the emergence of LRP, we focused our comparison of de-esterified homogalacturonan distribution on pericycle-endodermis junctions. Using an antibody that preferentially binds low-esterified homogalacturonan (*36*), immunogold labeling (*37*, *38*) followed by transmission electron microscopy identified two types of gold particle distribution in three-way pericycle-endodermis junctions (junctions of two pericycle cells and one endodermis cell or one pericycle cell and two endodermis cells). *gnom*^{184} junctions were more particle dense compared with WT junctions (Fig. 4, B to D, and fig. S7, B to E), which may prevent LR emergence. A more qualitative analysis showed a similar trend (fig. S7A). This indicates that GNOM mediates homogalacturonan distribution in pericycle-endodermis junctions.

Homogalacturonan subtypes are differentially distributed at sites of LRP initiation

We next investigated whether we could identify a direct link between homogalacturonan distribution and LRP inception. Using immunohistochemistry on transverse sections at the site of LRP inception, we found that pericycle-endodermis and endodermis-cortex circumference boundaries adjacent to the LRP have reduced levels of de-esterified homogalacturonan compared with the opposite (control) side (Fig. 4, E and G). By contrast, we identified higher levels of high-esterified homogalacturonan in the endodermis-cortex boundary (but not in the pericycle-endodermis) adjacent to the LRP compared with the opposite (control) side (Fig. 4, F and H). These results show that de-esterified homogalacturonan, which is usually associated with increased cell wall stiffness, is reduced near emerging LRP, and that pectin subtypes are differentially distributed at sites of initiating LRP. Given that LRP emerge by pushing through overlying cell layers, loss of de-esterified pectin could reduce cell adhesion at prebranch sites, providing a competence to form LRP.

Discussion

A minimum requirement for an oscillating system is delayed negative feedback, which can produce a periodic pattern with as few as two components (*39*). However, the somitogenesis clock requires the function of multiple components arranged in three complex networks (*40*). In *Arabidopsis*, more than 30 genes have been implicated in LR development, most of which are related to auxin signaling, transport, and homeostasis (*41*). A recent study showed that endodermis cells overlying the pericycle undergo volume reduction, which is required for the first cell division during LRP initiation (*6*). Although the precise mechanism by which pectin modification modulates the root clock and LRP initiation is still not known, we show here that LRP initiation involves differential distribution of pectin species, with reduced de-esterified pectin levels in cells overlying primordia. This distribution could allow the physical emergence of LRP while preventing emergence on the opposite side of the root. Additional support for the hypothesis that a balanced esterification level is essential for the root clock and LRP initiation is demonstrated by overexpression of either PME5 or PMEI3, which confers similar phenotypes in the root (this study) and in pavement cells (*42*). The fact that two proteins (GNOM and AGD3) with opposing functionalities also have an opposite phenotype provides additional evidence for the importance of balanced pectin components. One of the remaining questions is how the status of extracellular pectin is conveyed to the cell to initiate LR formation. One possibility is that biomechanical changes in the cell wall are translated into signal transduction that leads to LRP initiation. Alternatively, pectin might bind and activate a membranous receptor such as the FERONIA crRLK to initiate a signaling cascade that eventually results in the formation of LRP.

REFERENCES AND NOTES

1. M. A. Moreno-Risueno *et al.*, *Science* **329**, 1306–1311 (2010).
2. I. De Smet *et al.*, *Development* **134**, 681–690 (2007).
3. H. Fukaki, S. Tameda, H. Masuda, M. Tasaka, *Plant J.* **29**, 153–168 (2002).
4. J. C. Wilmoth *et al.*, *Plant J.* **43**, 118–130 (2005).
5. W. Xuan *et al.*, *Science* **351**, 384–387 (2016).
6. J. E. M. Vermeer *et al.*, *Science* **343**, 178–183 (2014).
7. K. Okumura *et al.*, *Plant Cell Physiol.* **54**, 406–417 (2013).
8. S. Naramoto *et al.*, *Plant Cell* **26**, 3062–3076 (2014).
9. D. E. Shevell *et al.*, *Cell* **77**, 1051–1062 (1994).
10. S.-J. Kim, F. Brandizzi, *Glycobiology* **26**, 940–949 (2016).
11. S. Picelli *et al.*, *Nat. Protoc.* **9**, 171–181 (2014).
12. F. Micheli, *Trends Plant Sci.* **6**, 414–419 (2001).
13. W. G. Willats *et al.*, *J. Biol. Chem.* **276**, 19404–19413 (2001).
14. R. Louvet *et al.*, *Planta* **224**, 782–791 (2006).
15. S. Wolf, T. Rausch, S. Greiner, *Plant J.* **58**, 361–375 (2009).
16. K. Tatematsu *et al.*, *Plant Cell* **16**, 379–393 (2004).
17. B. De Rybel *et al.*, *Curr. Biol.* **20**, 1697–1706 (2010).
18. K. Swarup *et al.*, *Nat. Cell Biol.* **10**, 946–954 (2008).
19. J. M. Alvarez *et al.*, *Plant J.* **80**, 1–13 (2014).
20. J. E. Lima, S. Kojima, H. Takahashi, N. von Wirén, *Plant Cell* **22**, 3621–3633 (2010).
21. D. Loqué, U. Ludewig, L. Yuan, N. von Wirén, *Plant Physiol.* **137**, 671–680 (2005).
22. H. Reinhardt *et al.*, *Plant Physiol.* **170**, 1640–1654 (2016).
23. W. Lin, W. Tang, C. T. Anderson, Z. Yang, FERONIA's sensing of cell wall pectin activates ROP GTPase signaling in Arabidopsis. bioRxiv 269647 [Preprint]. 22 February 2018. https://doi.org.10.1101/269647.
24. A. Peaucelle *et al.*, *Curr. Biol.* **18**, 1943–1948 (2008).
25. G. Wachsman, J. L. Modliszewski, M. Valdes, P. N. Benfey, *Plant Physiol.* **174**, 1307–1313 (2017).
26. R. Siligato *et al.*, *Plant Physiol.* **170**, 627–641 (2016).
27. N. Geldner *et al.*, *Cell* **112**, 219–230 (2003).
28. G. Wu *et al.*, *Sci. Signal.* **4**, ra29 (2011).
29. K. Koizumi *et al.*, *Development* **132**, 1699–1711 (2005).
30. J. G. Donaldson, C. L. Jackson, *Nat. Rev. Mol. Cell Biol.* **12**, 362–375 (2011).
31. Y. Rui *et al.*, *Plant Cell* **29**, 2413–2432 (2017).
32. F. Liners, J.-J. Letesson, C. Didembourg, P. Van Cutsem, *Plant Physiol.* **91**, 1419–1424 (1989).
33. Y. Verhertbruggen, S. E. Marcus, A. Haeger, J. J. Ordaz-Ortiz, J. P. Knox, *Carbohydr. Res.* **344**, 1858–1862 (2009).
34. Y. Verhertbruggen, J. L. Walker, F. Guillon, H. V. Scheller, *Front. Plant Sci.* **8**, 1505 (2017).
35. D. E. Shevell, T. Kunkel, N. H. Chua, *Plant Cell* **12**, 2047–2060 (2000).
36. S. E. Marcus *et al.*, *Plant J.* **64**, 191–203 (2010).
37. K. Ruel, Y. Nishiyama, J.-P. Joseleau, *Plant Sci.* **193-194**, 48–61 (2012).
38. J. K. Polko *et al.*, *Curr. Biol.* **28**, 3174–3182.e6 (2018).
39. A. Gierer, H. Meinhardt, *Kybernetik* **12**, 30–39 (1972).
40. A. Goldbeter, O. Pourquié, *J. Theor. Biol.* **252**, 574–585 (2008).
41. B. Péret *et al.*, *Trends Plant Sci.* **14**, 399–408 (2009).
42. K. T. Haas, R. Wightman, E. M. Meyerowitz, A. Peaucelle, *Science* **367**, 1003–1007 (2020).

ACKNOWLEDGMENTS

We thank C. Wilson and Y. Rui for technical support and advice; A. Peaucelle for seeds; M. Plue (Duke, SMif), R. Vancini and H. Mekeel (Duke Pathology), and K. Pryer for assistance with immunotissue processing and TEM microscopy; the Duke Center for Genomic and Computational Biology; and I. W. Taylor, R. Shahan, J. Vermeer, and J. K. Polko for critical review of the manuscript. **Funding:** This work was supported by grants to P.N.B. from the National Institutes of Health (grant nos. R01-GM043778 and R35-GM131725), the Howard Hughes Medical Institute, and the Gordon and Betty Moore Foundation (grant no. GBMF3405). M.A.M.-R. was funded by Ministerio de Economía y Competitividad de Spain (MINECO) and ERDF (grant no. BFU2016-80315-P). **Author contributions:** G.W. and P.N.B. conceived the project. G.W. and J.Z. performed the experiments. M.A.M.-R. assisted with mapping. C.T.A. provided supervision and resources for immunohistochemistry. G.W. wrote an initial draft of the manuscript. P.N.B., J.Z., M.A.M.-R., and C.T.A. reviewed and edited the manuscript. **Competing interests:** The authors declare no competing interests. P.N.B. is the cofounder and chair of the scientific advisory board of Hi Fidelity Genetics, Inc., a company that works on crop root growth. **Data and materials availability:** All materials are available upon request. All data are available in the main paper or the supplementary materials. The RNA-seq and DNA-seq datasets are available in the Sequence Read Archive repository with the accession numbers PRJNA478564 and PRJNA490692, respectively.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/819/suppl/DC1
Materials and Methods
Figs. S1 to S7
Table S1 to S5
References
Movie S1
Datasets S1 to S3
Code S1 to S9
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

15 March 2020; accepted 1 October 2020
10.1126/science.abb7250

MARTIAN ATMOSPHERE

Hydrogen escape from Mars is driven by seasonal and dust storm transport of water

Shane W. Stone[1]*, Roger V. Yelle[1], Mehdi Benna[2,3], Daniel Y. Lo[1], Meredith K. Elrod[2,4], Paul R. Mahaffy[5]

Mars has lost most of its once-abundant water to space, leaving the planet cold and dry. In standard models, molecular hydrogen produced from water in the lower atmosphere diffuses into the upper atmosphere where it is dissociated, producing atomic hydrogen, which is lost. Using observations from the Neutral Gas and Ion Mass Spectrometer on the Mars Atmosphere and Volatile Evolution spacecraft, we demonstrate that water is instead transported directly to the upper atmosphere, then dissociated by ions to produce atomic hydrogen. The water abundance in the upper atmosphere varied seasonally, peaking in southern summer, and surged during dust storms, including the 2018 global dust storm. We calculate that this transport of water dominates the present-day loss of atomic hydrogen to space and influenced the evolution of Mars' climate.

The enrichment of D/H in the martian atmosphere and surface, relative to Earth, indicates that most of Mars' initial reservoir of H_2O has been lost to space (*1*–*4*). The standard scenario to explain this escape is that H_2O is confined to low altitudes below a hygropause (a cold layer at 40- to 50-km altitude at which H_2O condenses), where photodissociation and HO_x chemistry produce H_2. The H_2 then diffuses through the hygropause into the upper atmosphere where it is destroyed, producing H, which escapes to space (*2*, *5*–*7*). This model would produce steady H_2 diffusion and H escape. Observations have shown rapid order-of-magnitude variations in the abundance of H in the exosphere, the outermost region of the atmosphere, indicating that H escape varies seasonally and during dust storms (*8*–*10*). A possible cause for rapid variation in the exospheric H abundance and escape rate is direct transport of H_2O—the most abundant hydrogen-bearing species in the martian atmosphere (*2*)—into the middle and upper atmosphere, where it could be destroyed by reactions with ions to produce H (*11*). This transport would require a weakening of the hygropause, which could be warmed and raised in altitude as a result of increased solar irradiation around perihelion and/or heating caused by dust in the atmosphere. The middle-atmospheric (~15- to 90-km altitude) H_2O abundance is often higher than that implied by the equilibrium vapor pressure at the hygropause temperature (*12*, *13*) and varies seasonally (*14*, *15*), and transient events substantially alter the vertical distribution of H_2O (*14*, *16*, *17*).

We test this hypothesis using data from the Neutral Gas and Ion Mass Spectrometer (NGIMS) (*18*) on the Mars Atmosphere and Volatile EvolutioN (MAVEN) spacecraft (*19*). NGIMS samples neutral and ionic species in situ as MAVEN flies through the martian upper atmosphere. We search for the chemical intermediates that lie between H_2O delivered from the lower atmosphere and H escape from the top of the atmosphere; for example, H_2O^+ and H_3O^+.

[1]Lunar and Planetary Laboratory, University of Arizona, Tucson, AZ 85711, USA. [2]Planetary Environments Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA. [3]Center for Research and Exploration in Space Science and Technology, University of Maryland Baltimore County, Baltimore, MD 21250, USA. [4]Center for Research and Exploration in Space Science and Technology, University of Maryland College Park, College Park, MD 20742, USA. [5]Solar System Exploration Division, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA.
*Corresponding author. Email: stone@lpl.arizona.edu

In situ MAVEN measurements

The position of MAVEN's periapsis (i.e., closest approach to Mars) varied over the course of the mission, as shown in Fig. 1. Regional dust storms occurred during Mars years (MYs) 32 and 33, and a global dust storm occurred in MY 34, also shown in Fig. 1. MAVEN instruments directly sample the atmosphere during the periapse segment of its 4.5-hour orbit, in which periapsis occurs at roughly 150-km altitude (*19*, *20*). During the observations of the last regional dust storm of MY 32, MAVEN's periapsis precessed from 18°N to 8°N latitude, with measurements at the onset of the dust storm obtained near 13°N. This regional storm arose in the southern hemisphere near martian solar longitude [Mars-Sun angle (L_S)] 313°, well after perihelion (L_S = 251°) and southern summer solstice (L_S = 270°). Amid the global dust storm of MY 34, MAVEN's periapsis precessed from 21°S to 17°S latitude, with measurements at the onset of the dust storm obtained at ~18°S. This global dust storm occurred around L_S = 189°, before perihelion and just after southern spring equinox (L_S = 180°). Figure 1A shows the total column dust optical depth from a combination of instruments: the Thermal Emission Spectrometer on Mars Global Surveyor, the Thermal Emission Imaging System on Mars Odyssey, and the Mars Climate Sounder (MCS) on Mars Reconnaissance Orbiter (*21*, *22*). Figure 1B shows the dayside temperature in the middle atmosphere at the 50-Pa pressure level measured by MCS (*23*, *24*). These quantities indicate the onset, evolution, and distribution in latitude of the dust events and the atmospheric response.

Table 1. H_2O budget and H production. Two models with low and high H_2O are listed, representing periods between and during dust storms, respectively (*32*). The H_2O mixing ratios at 80 km for the low- and high-H_2O cases are 2 and 430 ppm. Total H production from H_2O is twice the difference between all H_2O destruction and all H_2O production. Total H production from H_2 is twice the rate of destruction of H_2 by CO_2^+.

Reaction	Column rate (cm^{-2} s^{-1})	
	Low H_2O	High H_2O
H_2O production		
$H_3O^+ + e^- \rightarrow H_2O + H$	0.1×10^7	1.99×10^8
H_2O destruction		
$CO_2^+ + H_2O \rightarrow H_2O^+ + CO_2$	1.6×10^7	6.35×10^8
$HCO^+ + H_2O \rightarrow H_3O^+ + CO$	0.5×10^7	7.56×10^8
$CO_2^+ + H_2O \rightarrow HCO_2^+ + OH$	0.5×10^7	2.18×10^8
H production		
Total H production from H_2O	5.0×10^7	2.85×10^9
Total H production from H_2	1.9×10^8	1.9×10^8

Figure 2 shows the variation of the abundance in the ionosphere (the charged portion of the atmosphere) of H_2O^+ and H_3O^+, normalized by the abundance of electrons, near MAVEN's periapsis over the course of the mission. These two species are produced mainly by charge transfer between CO_2^+ and H_2O, producing H_2O^+, and proton transfer from HCO^+ to H_2O, producing H_3O^+ (see below). Three forms of variations are visible in Fig. 2A. First, there is diurnal variation across two orders of magnitude. The solar zenith angle (the angle between the vertical and the Sun) in Fig. 2, A to C, corresponds to the time of day at which the measurement was made (0° is noon at the equator, 90° dawn or dusk, and 180° midnight). Second, H_2O^+ varies with season (L_S), with repeatable maxima near L_S = 270°, the peak of southern summer, and minima near L_S = 40°. The relative abundance of H_2O^+ in the dayside ionosphere varies by more than an order of magnitude over a martian year, from tens of parts per million (ppm) to hundreds of ppm, peaking in southern summer near perihelion and declining during northern summer. The peak in ratio of H_2O^+ abundance to electron abundance ($[H_2O^+]/[e^-]$) in southern summer coincides with the peak in observed seasonal trends in middle-atmospheric H_2O abundance (*14*, *15*), the exospheric H and D abundances

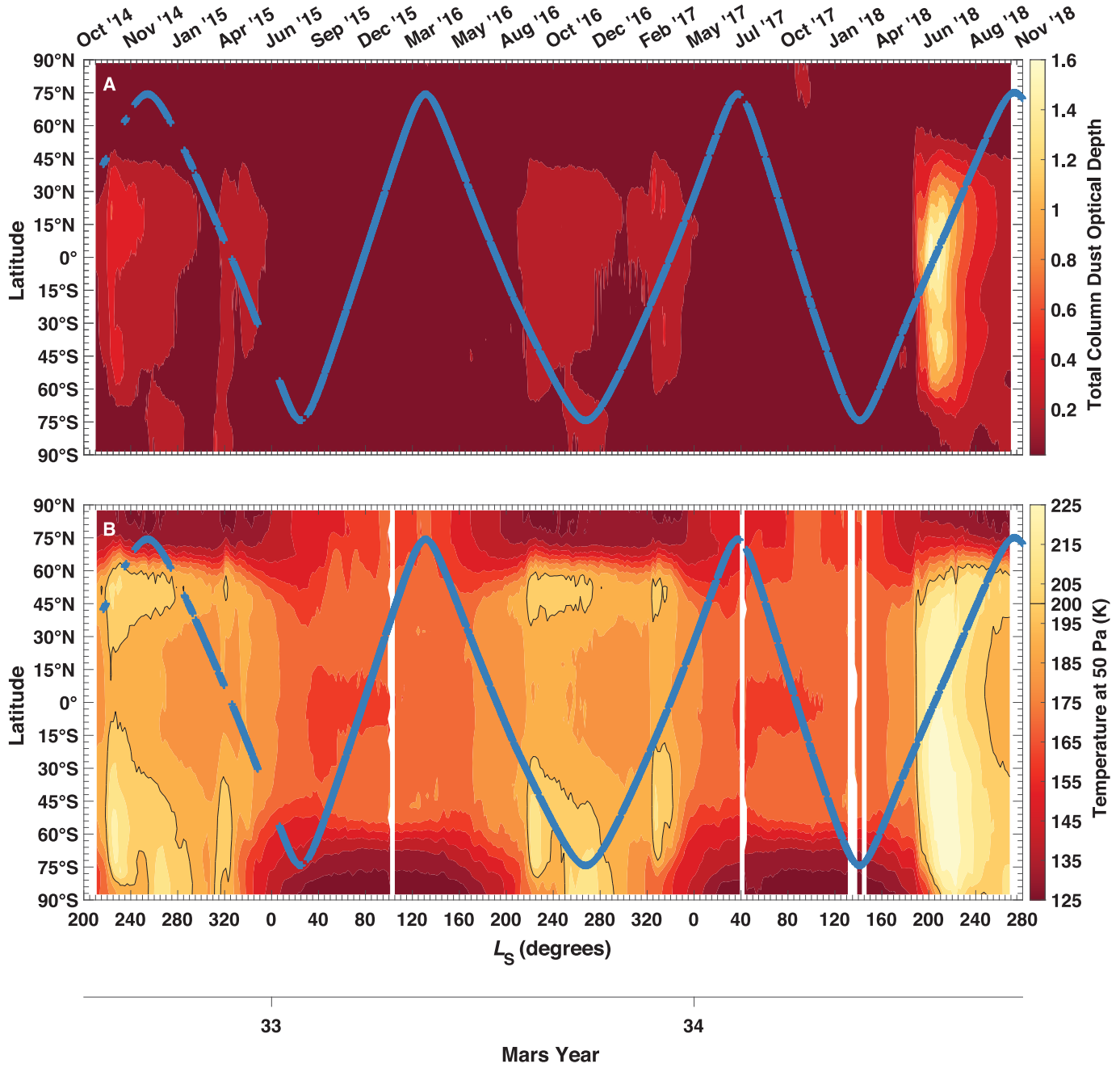


Fig. 1. MAVEN location and atmospheric conditions during martian dust storms. (**A**) Gridded total column dust optical depth (red to yellow) (*21*, *22*) and (**B**) mean dayside temperature (red to yellow with black contour at 200 K) at the 50-Pa pressure level from MCS measurements (*23*, *24*) between October 2014 and November 2018. MAVEN's periapsis positions are indicated with blue crosses. Gaps in MAVEN's periapsis location correspond to gaps in the NGIMS dataset (*32*). Ticks in the MY axis occur at the beginning of the year (L_S = 0°). Dust storms correspond with heating in the middle atmosphere.

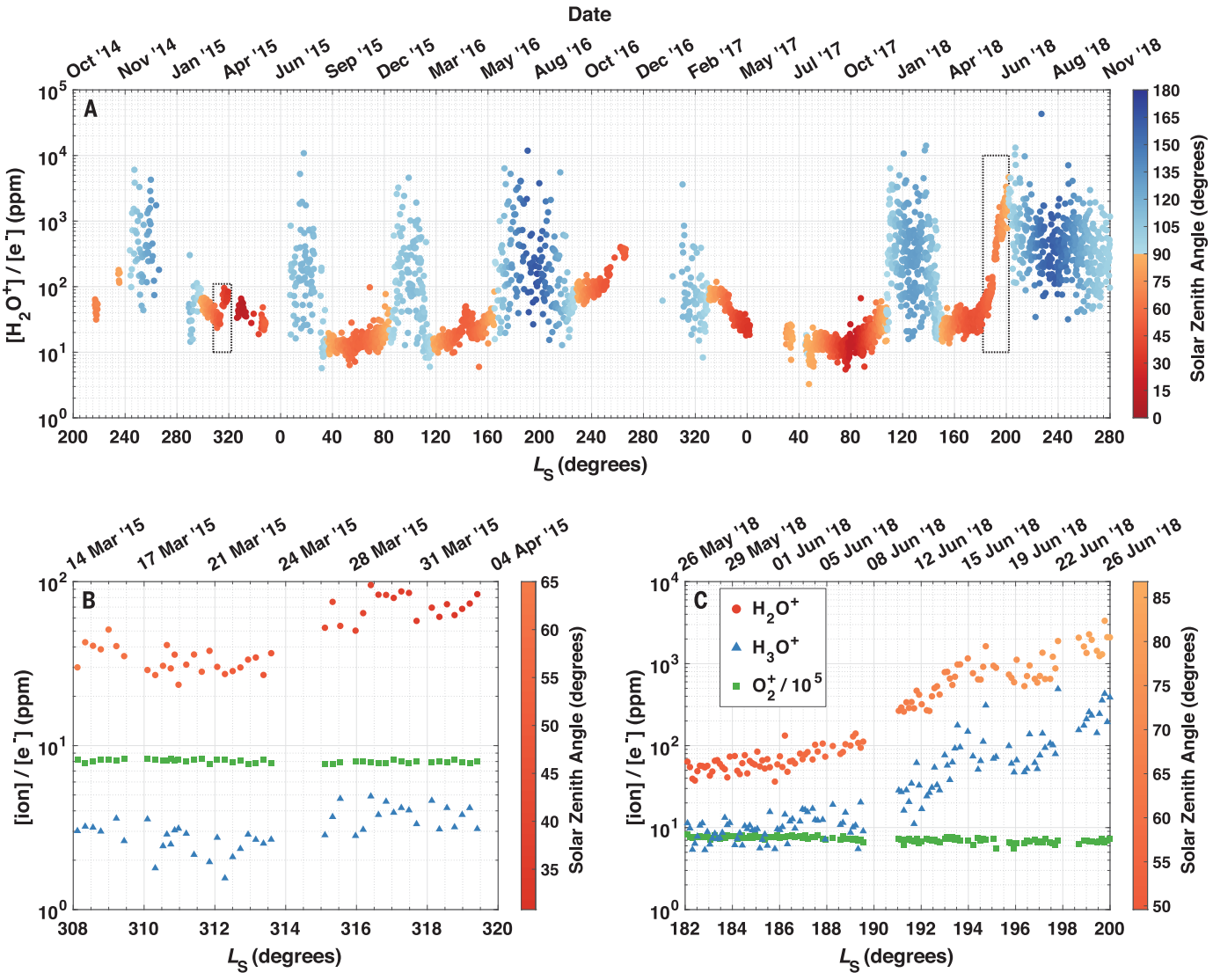


Fig. 2. Variation of H_2O ions measured using NGIMS. Mean $[H_2O^+]/[e^-]$ for each MAVEN orbit (**A**) between October 2014 and November 2018, (**B**) during a regional dust storm in MY 32, and (**C**) during the MY 34 global dust storm. Dotted boxes in (A) indicate the regions shown in (B) and (C). Each data point is the mean $[H_2O^+]/[e^-]$ (circles), $[H_3O^+]/[e^-]$ [blue triangles in (B) and (C)], or $[O_2^+]/[e^-]$ [green squares in (B) and (C)] between CO_2 densities of 5×10^8 to 10^9 cm^{-3} (MAVEN's nominal periapsis) for a single orbit. The value of the solar zenith angle is indicated by the color of the circle, as indicated in the colorbar. The transition from red to blue for $[H_2O^+]/[e^-]$ corresponds to the shift from day to night.

(*9*, *10*, *25*), and H escape rate (*26*–*28*). Third, Fig. 2A shows large and rapid increases in the H_2O^+ abundance coincident with dust storms near L_S = 315° in MY 32 and L_S = 190° in MY 34. More detailed views of these events are shown in Fig. 2, B and C.

The dust storms shown in Fig. 2, B and C, are substantial perturbations to the observed seasonal trend. At the onset in the upper atmosphere of the global dust storm of June 2018 (MY 34, Fig. 2C), the relative abundance of H_2O^+ in the ionosphere rises sharply in a short period of time: $[H_2O^+]/[e^-]$ increases from 106 to 327 ppm, a factor of 3.1, over ~2 days, from 8 June 2018 to 10 June 2018. Simultaneously, $[H_3O^+]/[e^-]$ increases from 11 to 28 ppm, a factor of 2.5. MAVEN's periapsis precessed onto the night side of Mars shortly after the onset of the global dust storm, and measurements in this period are affected by the diurnal variation of the ionosphere (see below). Nevertheless, there is an abrupt increase in the relative abundance of these ions, which coincides with the first effects of the dust storm on the upper atmosphere on 8 June 2018, as measured by the Imaging Ultraviolet Spectrograph on MAVEN (*29*) and neutral atmosphere density measurements from NGIMS (*30*). The regional dust storm observed in March 2015 (MY 32, Fig. 2B) occurs during the declining seasonal trend, with an upward perturbation caused by the storm: Over 2 days, $[H_2O^+]/[e^-]$ increases from 32 to 73 ppm, a factor of 2.3, and $[H_3O^+]/[e^-]$ increases from 2.4 to 3.9 ppm, a factor of 1.6.

Large diurnal variations are also observed in the relative abundance of H_2O^+ as MAVEN's periapsis repeatedly precessed from the day side to the night side, a process which occurs over several months, as seen in Fig. 2A. The relative abundance of H_2O^+ increases on the night side because, relative to other species in the martian atmosphere, H_2O has a low ionization potential (IP), 12.6 eV, and OH has a high proton affinity (PA), 6.14 eV. In general, atmospheric ionization flows by electron transfer from neutral species with lower IPs to positive ions whose parent neutrals have higher IPs and, by proton transfer, from species whose nonprotonated forms have lower PAs to species that have higher PAs (*31*). On the night side, away from a constant source

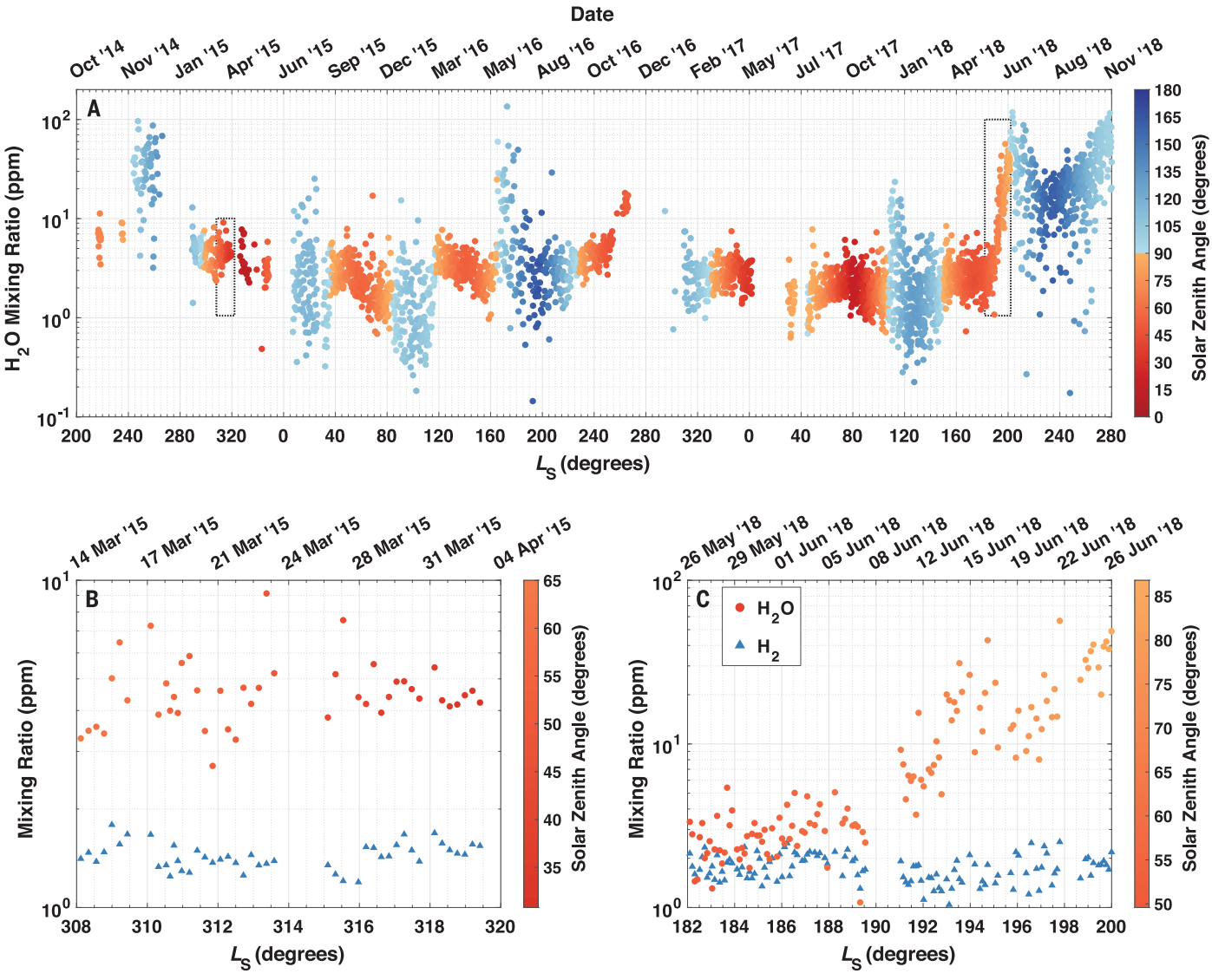


Fig. 3. Variation of H_2O and H_2. (**A** to **C**) Same as in Fig. 2 but for the H_2O mixing ratio (circles) and H_2 mixing ratio [blue triangles in (B) and (C)].

of photoionization driving the production of short-lived ions (such as O_2^+), H_2O^+ becomes more abundant. Thus, we can only observe the effects of dust storms on the day side; otherwise, the diurnal variation overwhelms the signal because of the storm.

Water in the upper atmosphere

Assuming that the increases in H_2O^+ and H_3O^+ are due to an influx of H_2O into the upper atmosphere, we calculate H_2O abundances from NGIMS density measurements of H_2O^+, H_3O^+, HCO^+, and e^- obtained over 4254 MAVEN orbits. We assume photochemical equilibrium for H_3O^+; i.e., its production rate is equal to its loss rate. This assumption is known to be valid in the thermosphere on the day side but has not been substantiated on the night side (*32*). H_3O^+ is mainly produced by proton transfer to H_2O and is only destroyed by dissociative recombination with e^- (*31*–*34*). The results of these calculations are shown in Fig. 3. The diurnal variation of H_2O is much smaller than that of H_2O^+, as expected in the neutral atmosphere, which is more chemically stable than the ionosphere.

A repeatable seasonal trend occurs in the H_2O abundance in the upper atmosphere, which peaks in southern summer between L_S = 250° and 270°, shown in Figs. 3A and 4. This trend in upper-atmospheric H_2O coincides with the seasonal trends seen in previous observations of the H_2O abundance in the middle atmosphere (*14*, *15*), the exospheric H and D abundances (*9*, *10*, *25*), and H escape flux (*26*–*28*). The mean upper-atmospheric H_2O mixing ratio is >1 ppm at all times and can exceed 10 ppm in southern summer. A sinusoidal model fitted to the dayside H_2O abundances from all MYs, but excluding regional and global dust events, is shown in Fig. 4. This model has a maximum H_2O mixing ratio of 4.3 ppm at L_S = 259° and a minimum of 2.0 ppm at L_S = 86°, for a variation of more than a factor of two. The H_2O mixing ratio is never negligible, contrary to the assumption of 0 to 4 parts per billion made in numerous modeling studies (*5*, *33*, *35*).

During the global dust storm of June 2018, the mean H_2O mixing ratio increased by a factor of 2.4, from a mean value of 3.0 to 7.1 ppm over 2 days. The water abundance then continued to increase following the seasonal and dust storm trends, reaching values >60 ppm at L_S = 204°. This event included the highest persistent H_2O abundances we observed, holding at tens of ppm for more than 5 months at the end of 2018. During the regional storm observed in MY 32, the H_2O mixing ratio increases by 7%, from 4.6 to 4.9 ppm in 2 days. Thus, dust storms drive perturbations in upper-atmospheric H_2O that are far more rapid than the diurnal variations, and even small dust

storms have a measurable impact. The MY 32 dust storm occurred when MAVEN was near the subsolar point and is superimposed on a declining seasonal trend. We interpret these variations as being due to enhanced transport of H_2O into the upper atmosphere during the dust storms.

Previous and coincident measurements of water

The effects of both regional and global dust storms on the H_2O in the lower and middle atmosphere have been observed before. In MY 32, Mars Express observed large seasonal increases in the H_2O mixing ratio, including an increase around L_S = 315° (*14*), coinciding with the increase we observe in Figs. 2 and 3. These measurements indicate an increase in the hygropause altitude and an injection of water into the middle atmosphere, coinciding with our observed increases in the abundances of H_2O, H_2O^+, and H_3O^+ in the upper atmosphere. Similarly, MCS detected an increase in the hygropause altitude, rising to 65 to 70 km, in this period (*15*). During the global dust storm of MY 34, Trace Gas Orbiter (TGO) measured a peak H_2O abundance of ~100 ppm at 50-km altitude in the southern hemisphere (80°S to 83°S) and a peak H_2O abundance of ~250 ppm just below 40-km altitude in the northern hemisphere (51°N to 59°N), indicating an injection of water into the middle atmosphere coincident with our observations (*16*). Both of these measurements were obtained at L_S = 196.64°. Further TGO measurements showed that H_2O abundances of ~5 to 10 ppm extended to altitudes up to 90 to 100 km (*13*, *17*). The time scales of dust storm–induced variations in the upper-atmospheric H_2O mixing ratio shown in Fig. 3 are consistent with these simultaneous measurements of the middle-atmospheric H_2O mixing ratio during the 2018 global storm. MCS observed an H_2O mixing ratio close to 300 ppm near 50-km altitude in the middle atmosphere (*36*). At the same time, we measure a mean H_2O mixing ratio of 7.5 ppm near MAVEN's periapsis (~150-km altitude) over an L_S range of 191.05° to 193.21° and a latitude range of 13°S to 18°S. These measurements in the middle atmosphere, below the NGIMS sampling range, provide context for the NGIMS measurements. Seasonally and during dust storms, the H_2O abundance in the middle atmosphere increased substantially in the span of a few martian days, coincident with NGIMS observations in the upper atmosphere.

Molecular hydrogen in the upper atmosphere

Molecular hydrogen, H_2, may contribute to variations in protonated ions and is a source of H in the upper atmosphere. Measured H_2 variations obtained by NGIMS are shown in Fig. 3, B and C. H_2 reacts with numerous ions, mainly CO_2^+, to produce H atoms that may escape the planet's atmosphere. H_2 itself can escape, although much more slowly than H because H_2 is twice the mass of H and lighter species have lower escape energies. The absolute sensitivity of NGIMS to H_2 is unknown, so we cannot derive an absolute H_2 abundance. However, relative variations of H_2 are measurable. No seasonal trend or impact of dust storms are apparent in the upper-atmospheric H_2 abundance (Fig. 3, B and C). H_2 is therefore not the source of H responsible for seasonal and dust storm–induced variations in the upper atmosphere.

Water as a source of escaping hydrogen

We constructed a photochemical model to investigate the effects of the H_2O abundances in the Mars ionosphere on the H escape rate. We constrain the model with data from a MAVEN Deep Dip (DD) campaign. During weeklong DD campaigns, the MAVEN spacecraft penetrates to lower altitudes, ~125 km. The measured altitude profiles provide constraints on ionospheric composition. We use the second DD campaign (DD2) executed around L_S = 329° in MY 32, because MAVEN's periapsis was then close to the subsolar point. We average the measurements over the DD using a previously described technique to remove wave structure in the altitude profiles (*20*). The solar zenith angles and latitudes do not vary appreciably during a DD, so the resulting profiles are effectively averages over longitude for a latitude of 4°S and local time of 12 p.m.

The photochemical calculations are one dimensional (1D) and use the DD2 geometry. Neutral and ionic composition is calculated from 80 to 250 km, including photolysis, chemical reactions, and diffusion of neutrals (*32*). We include both molecular and eddy diffusion, adopting an eddy diffusion coefficient of 10^8 cm^2 s^{-1}. Ions are assumed to be in local photochemical equilibrium. Rate coefficients are discussed below.

The model is constructed to match the primary neutral constituents measured by NGIMS, as shown in Fig. 5A. We calculate H_2O by assuming a mixing ratio of 2 ppm at 80 km, consistent with our measurements obtained near aphelion (L_S = 71°) in the absence of dust storms (Fig. 4). Photolysis causes the H_2O mixing ratio to decrease slightly with altitude until ~160 km, where molecular diffusion begins to dominate and the mixing ratio starts to increase with height. For O, we assume a mixing ratio of 2% at 80 km, chosen to produce a vertical profile consistent with NGIMS measurements at higher altitudes.

Figure 5B shows that the altitude profiles of H_2O^+ and H_3O^+ in the model are consistent

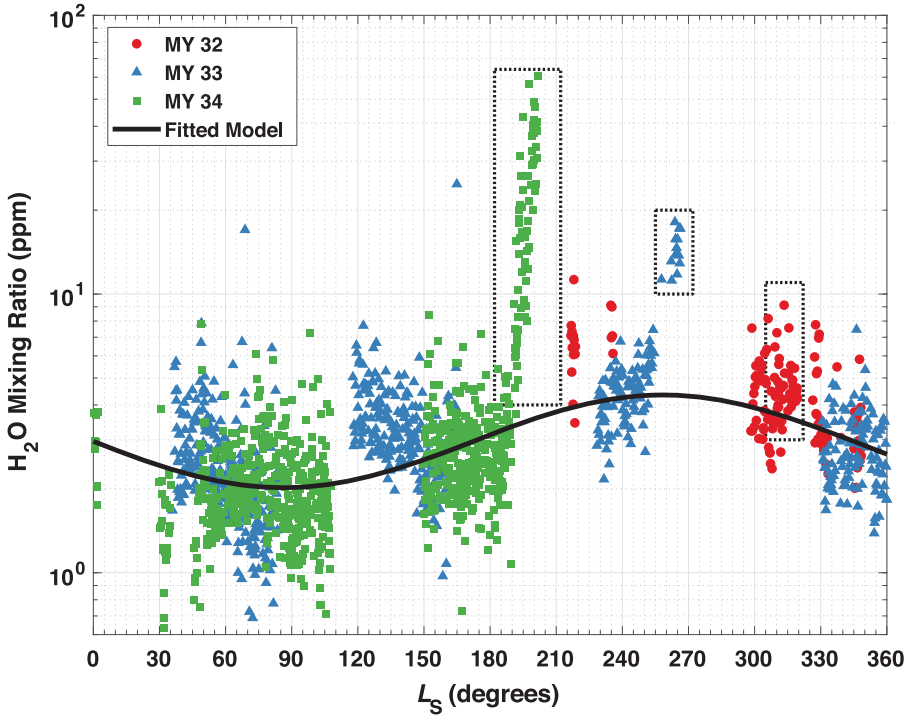


Fig. 4. Seasonal variation of upper-atmospheric H_2O. Each data point is the mean dayside (solar zenith angle <90°) H_2O mixing ratio between CO_2 densities of 5 × 10^8 and 10^9 cm^{-3} for a single MAVEN orbit over MYs 32 (red circles), 33 (blue triangles), and 34 (green squares). The solid line is a sinusoidal model fitted to the data from all three MYs to illustrate the seasonal variation. Dust storm data (dotted boxes) are included in the plot but excluded from the model fitting.

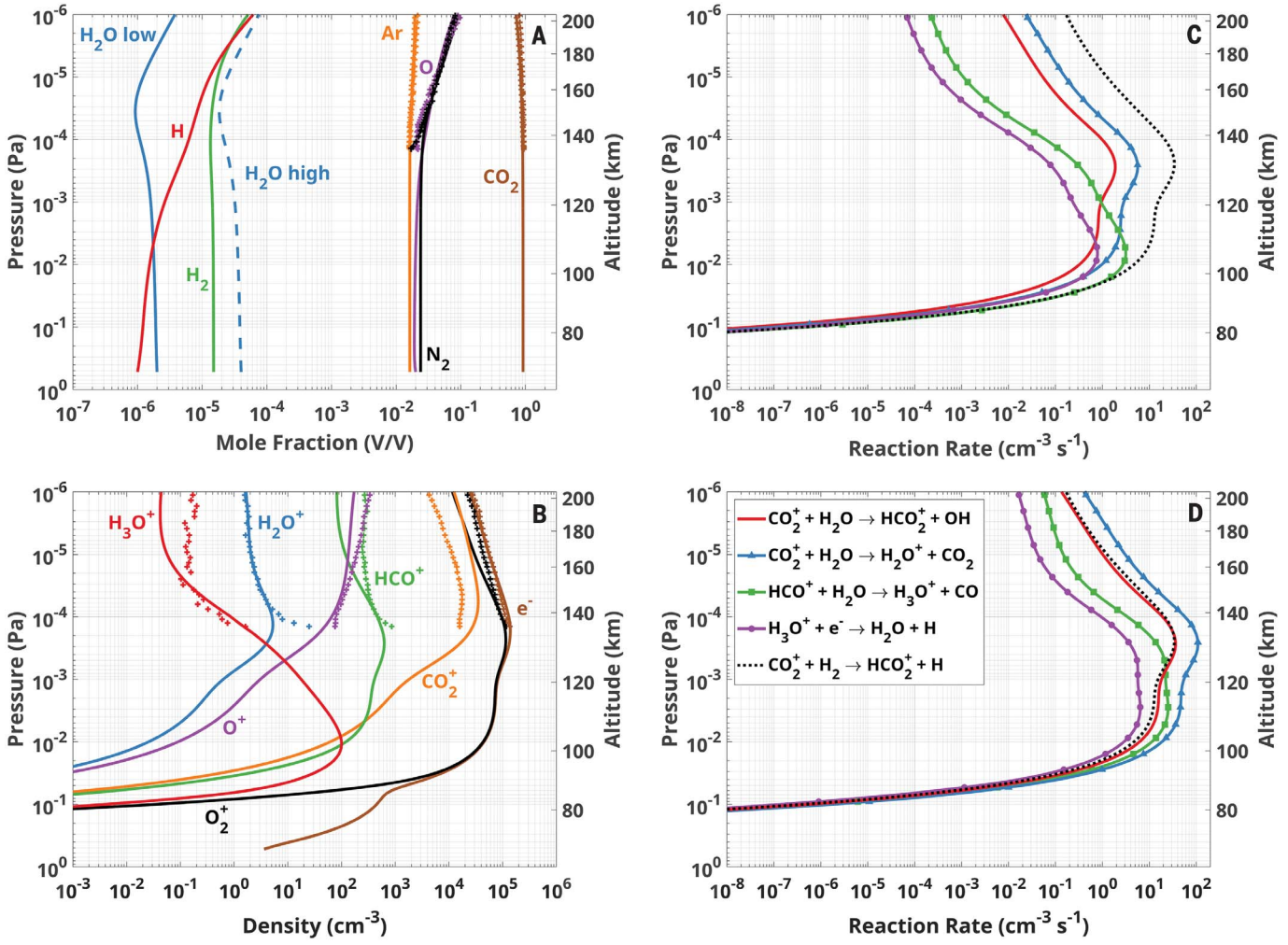


Fig. 5. Calculated 1D model abundances and reaction rates. (**A**) Neutral species abundances from our low-H_2O photochemical model (lines) compared with the NGIMS DD2 mean data (crosses). (**B**) Same as in (A), but for ionic species. Chemical reaction rates for selected reactions in the (**C**) low- and (**D**) high-H_2O models. The legend in (D) also applies to (C).

with the observations, validating our earlier inferences. The model also matches measurements of CO_2^+, O_2^+, and HCO^+, species involved in the chemical destruction of H_2O. Our calculations also match the O^+ observations. Models of the Mars ionosphere have often assumed that O^+ was produced by $CO_2^+ + O \rightarrow O^+ + CO_2$ with a rate coefficient of 10^{-10} cm^3 s^{-1} (*37*); however, laboratory work has placed an upper limit of 6×10^{-13} cm^3 s^{-1} on this rate coefficient (*38*, *39*). Adopting this value, O^+ is produced primarily by $CO_2 + h\nu \rightarrow CO + O^+$ ($h\nu$, photon energy) and lost by $O^+ + CO_2 \rightarrow CO_2^+ + O$ and $O^+ + CO_2 \rightarrow O_2^+ + CO$. Previous models assumed a combined rate coefficient for these reactions of 10^{-9} cm^3 s^{-1} (*40*). However, state-specific measurements indicate a value of 4×10^{-10} cm^3 s^{-1} (*41*). Our model uses a value of 3×10^{-10} cm^3 s^{-1}, chosen to match DD2 observations. Between 140 and 150 km, the predicted and observed protonated ion densities agree to within 25% (Fig. 5B). There are larger discrepancies at higher altitudes and near MAVEN's periapsis, but the observed densities near periapsis are uncertain because the spacecraft is moving primarily horizontally (*20*) and the model densities are uncertain at higher altitudes because ion diffusion has been neglected in the model.

We use the model to calculate the net destruction rate of H_2O and the production rate of H in the ionosphere (Table 1). The NGIMS measurements are only available above the ionospheric peak, the location of the maximum total ion density (Fig. 5B, 135 km), so we rely on model predictions to investigate chemical processes down to 80 km. In addition to the results for the DD2 model, we present rates for a similar model calculated for the same conditions and input data, but in which we adopt an H_2O mixing ratio of 430 ppm at 80 km, consistent with our measurements obtained during the global dust storm of MY 34 (Fig. 4). H production in the ionosphere for the low-H_2O model is 5×10^7 cm^{-2} s^{-1}, at the low end of the range of H escape rates inferred by previous studies, 10^7 to 5×10^9 cm^{-2} s^{-1} (*26*–*28*). The column-integrated H_2O destruction in the ionosphere for the high-H_2O model is 1.4×10^9 cm^{-2} s^{-1}, corresponding to an H production rate of 2.8×10^9 cm^{-2} s^{-1}, at the high end of the range determined in previous studies (*26*–*28*). H production from destruction of H_2 by CO_2^+ is 9.6×10^7 cm^{-2} s^{-1}, less than twice as large as H production from ionospheric destruction of H_2O in the low-H_2O models, and substantially less than that in the high-H_2O models.

It is possible that ionospheric destruction of H_2O is equally or more important than neutral photolysis of H_2O for H escape. We cannot compare the upper- and lower-atmosphere contributions to escape in a more quantitative way because escape rates in these models are closely dependent on the boundary conditions on the H and H_2 mixing ratios at 80 km,

which are assumed values chosen to replicate observations at higher altitudes. Boundary conditions on mixing ratios can introduce spurious fluxes into photochemical models. Physical boundary conditions, such as deposition velocities at the surface, are required to accurately calculate fluxes. Our calculations and the high H_2O densities in the thermosphere found in the MAVEN observations show that ionospheric chemistry is a source of H, which can escape from the Mars atmosphere.

Implications for climate evolution

Early studies of H production and escape neglected ionospheric destruction of H_2O because H_2O was assumed to be confined to low altitudes by a hygropause (*5*, *6*, *42*). Later, it was found that the hygropause varies in altitude with season, leading to speculation that H_2O saturation may not occur at all during dust storms, owing to elevated temperatures (*43*). More recent 1D photochemical models (*35*, *44*, *45*) and 3D global circulation models (*46*–*48*) have likewise not included chemical destruction of H_2O in the ionosphere. Occultation and limb profile measurements in the middle atmosphere demonstrated that the hygropause does not confine H_2O to low altitudes (*13*, *15*–*17*), and it has been postulated the elevated exospheric H densities during dust storms could be explained by these enhancements in the middle-atmospheric H_2O abundance (*11*). However, this model neglected ionospheric chemistry, assuming the H was produced solely by neutral photolysis. We have shown that H_2O is present in the ionosphere with substantial abundance throughout the martian year continuously since the arrival of MAVEN and that H is produced from this H_2O by ion chemistry.

The time scales for destruction of H_2O and production of H by ionospheric chemistry are short. Once in the ionosphere, H_2O is destroyed primarily by reactions with CO_2^+ with a rate coefficient of 2.4×10^{-9} cm^{-3} s^{-1} (*40*). Adopting a CO_2^+ density of 3×10^4 cm^{-3} (Fig. 5B, 135 km) implies a lifetime for H_2O of 4 hours. This is more than an order of magnitude shorter than the H_2O photolysis time constant of 2×10^5 s (*48*). Assuming an electron density of 10^5 cm^{-3} (Fig. 5B, 135 km) and electron recombination rates of 4.4×10^{-7} cm^{-3} s^{-1} (*49*) for H_2O^+ and H_3O^+ yields a time constant of ~20 s for production of H from electron recombination. Thus, H_2O transported to the ionosphere is quickly broken apart and converted to H. We expect the H produced at these high altitudes to have a high probability for escape. This implies that the limiting factor is the supply rate of H_2O to the ionosphere from lower altitudes. This has implications for our view of H escape from Mars. In previous models, H escape was controlled by production of H_2 in the lower atmosphere and its slow diffusion through the middle atmosphere. In our interpretation, H escape depends on how H_2O moves past the hygropause and the rate at which it is transported from the middle atmosphere to the ionosphere.

The relative importance of neutral and ion chemistry in the production and escape of thermal and nonthermal H remains unclear. Because we observe that the H_2 abundance is constant during the dust storms whereas the H_2O abundance increases strongly, ionospheric destruction drives the elevated H escape rate during dust events. The importance of ion chemistry to H production has been discussed previously (*34*), although without direct information on the H_2O abundance in the ionosphere. The H escape flux can be expressed as

$$F = 1.6 \times 10^8 + 1.4 \times 10^{13} f \text{ cm}^{-2} \text{ s}^{-1} \quad (1)$$

where f is the H_2O mixing ratio at 80 km (*34*). The constant term of 1.6×10^8 cm^{-2} s^{-1} is the limiting flux for an H_2 mixing ratio of 15 ppm, the value imposed at the lower boundary in those models (*34*). Adopting an H_2O mixing ratio of 2 ppm in Eq. 1 produces a contribution to escape from ion chemistry of 2.8×10^7 cm^{-2} s^{-1}, roughly half the column-integrated production rate in Table 1, suggesting an escape probability of 50% for the H produced in the ionosphere. This result is strongly dependent on the H mixing ratio at 80 km, which is an assumed value. Thus, in the previous model (*34*) and ours, the relative contributions of neutral and ion chemistry to escape depend on the boundary conditions for H and H_2 at 80 km. The abundance of, exact mechanism of production for, and contribution to escape of a nonthermal component of the H corona at Mars have not been determined because of degeneracies in the models used (*27*, *28*) and cannot be determined from our results.

NGIMS data are confined to altitudes above 125 km, whereas occultation and limb-sounding experiments have measured H_2O up to ~100 km, leaving a large gap in coverage between 100 and 125 km. Figure 5A shows our predicted H_2O mole fraction from the middle atmosphere through the ionosphere, but this is based on a 1D model, and the real atmosphere is likely to be more complex. The mixing ratios inferred for the middle atmosphere are often far greater than the values we calculate for the ionosphere (*13*, *16*, *17*). This contrast between the inhomogeneous H_2O distribution derived from the middle-atmosphere measurements and the roughly constant altitude profile predicted by the 1D model (Fig. 5) suggests that the middle-atmosphere measurements are sensing a local enhancement of H_2O that is transported horizontally around the planet as it diffuses upward. Such horizontal transport is not included in our 1D model. The largest peak in the lower-atmospheric H_2O mixing ratio occurs in northern summer (*2*), the period in which we observe a minimum in the upper-atmospheric H_2O mixing ratio (Fig. 4). The cause of this discrepancy between the seasonal variations in lower- and upper-atmosphere H_2O abundances is unknown and cannot be explained by a 1D model.

The seasonal and dust storm–mediated delivery of water to the upper atmosphere could have played a substantial role in the evolution of the martian climate from its warm and wet state billions of years ago to the cold and dry planet we observe today. Mars has likely lost enough H_2O to cover the planet's surface with an ocean tens to hundreds of meters deep, and loss rates must have been higher in the past (*4*). In a martian year with no dust storms, assuming the H_2O mixing ratio at 80 km is constant at 2 ppm, we calculate that 3.0×10^{15} cm^{-2} H would be produced from H_2O and 1.1×10^{16} cm^{-2} H by destruction of H_2 by CO_2^+, assuming all H, including that in HCO_2^+, is eventually liberated. During a global dust storm such as the one observed in MY 34, in which the H_2O mixing ratio at 80 km reaches 430 ppm, 1.1×10^{16} cm^{-2} H would be produced from H_2O in just 45 days. Thus, a single global dust storm can produce more than an entire martian year's worth of H production and escape. If we extrapolate these numbers into the past, we find that over a billion years, a 44-cm-deep global layer of H_2O would be lost owing to destruction of H_2O in the ionosphere, and, assuming a rate of about one global dust storm every 10 years (*50*), an additional 17 cm of H_2O would be lost owing to global dust storms. Larger loss estimates would be obtained if this extrapolation included the seasonal fluctuation of H_2O in the upper atmosphere (Fig. 4) and regional dust storms, which, relative to global dust storms, likely have a smaller but non-negligible impact on H escape. These effects, though, are smaller than uncertainties related to our lack of knowledge of the martian climate billions of years ago (*4*). H escape driven by direct transport of H_2O to the upper atmosphere would have accelerated the loss of water from Mars as the martian hygropause weakened owing to a decrease in atmospheric pressure, itself driven by atmospheric escape and the development of a dry, dusty surface.

REFERENCES AND NOTES

1. R. D. Wordsworth, *Annu. Rev. Earth Planet. Sci.* **44**, 381–408 (2016).
2. R. M. Haberle, R. T. Clancy, F. Forget, M. D. Smith, R. W. Zurek, Eds., *The Atmosphere and Climate of Mars* (Cambridge Univ. Press, 2017).
3. R. M. Ramirez, R. A. Craddock, *Nat. Geosci.* **11**, 230–237 (2018).
4. B. M. Jakosky *et al.*, *Icarus* **315**, 146–157 (2018).
5. D. M. Hunten, M. B. McElroy, *J. Geophys. Res.* **75**, 5989–6001 (1970).
6. T. D. Parkinson, D. M. Hunten, *J. Atmos. Sci.* **29**, 1380–1390 (1972).

7. M. B. McElroy, T. M. Donahue, *Science* **177**, 986–988 (1972).
8. M. S. Chaffin *et al*., *Geophys. Res. Lett.* **41**, 314–320 (2014).
9. D. Bhattacharyya, J. T. Clarke, J.-L. Bertaux, J.-Y. Chaufray, M. Mayyasi, *Geophys. Res. Lett.* **42**, 8678–8685 (2015).
10. J. T. Clarke *et al*., *J. Geophys. Res. Space Phys.* **122**, 2336–2344 (2017).
11. M. S. Chaffin, J. Deighan, N. M. Schneider, A. I. F. Stewart, *Nat. Geosci.* **10**, 174–178 (2017).
12. L. Maltagliati *et al*., *Science* **333**, 1868–1871 (2011).
13. A. A. Fedorova *et al*., *Science* **367**, 297–300 (2020).
14. A. A. Fedorova *et al*., *Icarus* **300**, 440–457 (2018).
15. N. G. Heavens *et al*., *New Astron.* **2**, 126–132 (2018).
16. A. C. Vandaele *et al*., *Nature* **568**, 521–525 (2019).
17. S. Aoki *et al*., *J. Geophys. Res. Planets* **124**, 3482–3497 (2019).
18. P. R. Mahaffy *et al*., *Space Sci. Rev.* **195**, 49–73 (2015).
19. B. M. Jakosky *et al*., *Space Sci. Rev.* **195**, 3–48 (2015).
20. S. W. Stone, R. V. Yelle, M. Benna, M. K. Elrod, P. R. Mahaffy, *J. Geophys. Res. Planets* **123**, 2842–2867 (2018).
21. L. Montabone *et al*., *Icarus* **251**, 65–95 (2015).
22. L. Montabone *et al*., *J. Geophys. Res. Planets* **125**, e2019JE006111 (2020).
23. A. Kleinböhl *et al*., *J. Geophys. Res. Planets* **114**, E10006 (2009).
24. D. M. Kass *et al*., *Geophys. Res. Lett.* **46**, 2019GL083931 (2019).
25. M. Mayyasi *et al*., *J. Geophys. Res. Space Phys.* **124**, 2152–2164 (2019).
26. J. S. Halekas, *J. Geophys. Res. Planets* **122**, 901–911 (2017).
27. D. Bhattacharyya *et al*., *J. Geophys. Res. Space Phys.* **122**, 11756–11764 (2017).
28. M. S. Chaffin *et al*., *J. Geophys. Res. Planets* **123**, 2192–2210 (2018).
29. J.-Y. Chaufray *et al*., *Geophys. Res. Lett.* **47**, e2019GL082889 (2020).
30. M. K. Elrod, S. W. Bougher, K. Roeten, R. Sharrar, J. Murphy, *Geophys. Res. Lett.* **47**, e2019GL084378 (2020).
31. J. L. Fox, *Icarus* **252**, 366–392 (2015).
32. Materials and methods are available as supplementary materials.
33. J. L. Fox, M. Benna, P. R. Mahaffy, B. M. Jakosky, *Geophys. Res. Lett.* **42**, 8977–8985 (2015).
34. V. A. Krasnopolsky, *Icarus* **321**, 62–70 (2019).
35. V. A. Krasnopolsky, *J. Geophys. Res. Planets* **107**, 5128 (2002).
36. N. G. Heavens, D. M. Kass, J. H. Shirley, *J. Geophys. Res. Planets* **124**, 2863–2892 (2019).
37. F. C. Fehsenfeld, D. B. Dunkin, E. E. Ferguson, *Planet. Space Sci.* **18**, 1267–1269 (1970).
38. J. L. Fox, A. S. Johnson, S. G. Ard, N. S. Shuman, A. A. Viggiano, *Geophys. Res. Lett.* **44**, 8099–8106 (2017).
39. J. E. Tenewitz *et al*., *J. Chem. Phys.* **148**, 084305 (2018).
40. V. G. Anicich, *J. Phys. Chem. Ref. Data* **22**, 1469–1569 (1993).
41. C. Alcaraz, C. Nicolas, R. Thissen, J. Zabka, O. Dutuit, *J. Phys. Chem. A* **108**, 9998–10009 (2004).
42. M. B. McElroy, *Science* **175**, 443–445 (1972).
43. B. M. Jakosky, *Space Sci. Rev.* **41**, 131–200 (1985).
44. H. Nair, M. Allen, A. D. Anbar, Y. L. Yung, R. T. Clancy, *Icarus* **111**, 124–150 (1994).
45. J. L. Fox, A. B. Hać, *Icarus* **204**, 527–544 (2009).
46. F. González-Galindo *et al*., *J. Geophys. Res. Planets* **120**, 2020–2035 (2015).
47. J.-Y. Chaufray *et al*., *Icarus* **245**, 282–294 (2015).
48. F. Daerden *et al*., *Icarus* **326**, 197–224 (2019).
49. M. J. Jensen *et al*., *Astrophys. J.* **543**, 764–774 (2000).
50. J. H. Shirley, C. E. Newman, M. A. Mischna, M. I. Richardson, *Icarus* **317**, 197–208 (2019).
51. S. W. Stone *et al*., Hydrogen Escape from Mars is Driven by Seasonal and Dust Storm Transport of Water, Version 2, Zenodo (2020); doi: 10.5281/zenodo.4010623.

ACKNOWLEDGMENTS

We acknowledge the contributions of the NASA MAVEN teams responsible for designing, constructing, and operating the spacecraft and its instruments, without which this work would be impossible. **Funding:** All authors were supported by funding from the MAVEN mission, part of the NASA Mars Exploration Program. **Author contributions:** S.W.S. and R.V.Y. conceived the study and wrote the draft manuscript. S.W.S., R.V.Y., M.B., M.K.E., and P.R.M. collected, calibrated, and analyzed the NGIMS data. R.V.Y. and D.Y.L. constructed the 1D photochemical model and analyzed the output, with the assistance of S.W.S. All authors contributed to the interpretation of results and preparation of the manuscript. **Competing interests:** We declare no competing interests. **Data and materials availability:** MAVEN data are available on the NASA Planetary Data System at https://atmos.nmsu.edu/data_and_services/atmospheres_data/MAVEN/maven_main.html and the MAVEN Science Data Center at https://lasp.colorado.edu/maven/sdc/public/. We used NGIMS Level 2, version 8, revision 1 data (*32*), which can be searched at https://lasp.colorado.edu/maven/sdc/public/pages/search/search.html. The 1D photochemical model code and output files are available at Zenodo (*51*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/824/suppl/DC1
Materials and Methods
Fig. S1
Table S1
References (*52*–*63*)

11 December 2019; accepted 11 September 2020
10.1126/science.aba5229

REPORT

METALLURGY

Constrained minimal-interface structures in polycrystalline copper with extremely fine grains

X. Y. Li[1]*, Z. H. Jin[1,2], X. Zhou[1], K. Lu[1]*

Metals usually exist in the form of polycrystalline solids, which are thermodynamically unstable because of the presence of disordered grain boundaries. Grain boundaries tend to be eliminated through coarsening when heated or by transforming into metastable amorphous states when the grains are small enough. Through experiments and molecular dynamics simulations, we discovered a different type of metastable state for extremely fine-grained polycrystalline pure copper. After we reduced grain sizes to a few nanometers with straining, the grain boundaries in the polycrystals evolved into three-dimensional minimal-interface structures constrained by twin boundary networks. This polycrystalline structure that underlies what we call a Schwarz crystal is stable against grain coarsening, even when close to the equilibrium melting point. The polycrystalline samples also exhibit a strength in the vicinity of the theoretical value.

The end-member structures for solids are single crystals, where atoms are arranged in ordered crystal lattices throughout the sample, and amorphous solids or glasses, where atomic arrangements are disordered with short- or medium-range ordering only (*1*). Metals usually exist in the form of polycrystalline solids, which are between these two extremes. They are made up of smaller crystallites (called grains) that are separated by a variety of boundaries where atomic arrangements are normally disordered. The disordered grain boundary (GB) is what makes polycrystalline metals thermodynamically unstable.

Polycrystalline materials tend to become more stable when GBs are eliminated, until they ultimately become single crystals. An example of this process is how the grains in polycrystals may coarsen through GB migration that usually occurs at temperatures below half of the melting point (*2*). The coarsening temperature drops with decreasing grain size, going as far down as room temperature in some nanograined metals (*3*). For fine-grained polycrystals with a high enough GB density, transforming into a metastable amorphous state is an alternative option to stabilization. This state is anticipated from a thermodynamic point of view, and amorphization has been observed in a number of metallic alloys [e.g., Ni-P (*4*) and Ni-W (*5*)] as grain sizes were reduced below certain limits (typically a few nanometers). However, under the thermodynamic and kinetic constraints, amorphous states are compositionally selective and rarely form for most metallic alloys and pure metals under conventional conditions. A fundamental question to be answered is whether other metastable structures may be adopted when polycrystalline grains are steadily refined to extremely small scales.

We have recently found that when grains of pure Cu and Ni are refined to a few tens of nanometers in size through plastic deformation, the autonomous GB relaxation into low-energy states is triggered with GB dissociations (*6*). The corresponding GB energy decreases accordingly, which leads to a substantial elevation in the thermal and mechanical stabilities of nanograins against coarsening at smaller sizes (*6*, *7*). This observation suggests the possibility that nanograined structures may evolve into more-stable states by approaching the grain-size extreme. We discovered, through experimental and molecular dynamics (MD) simulations, a metastable state in polycrystalline pure Cu with grain sizes of a few nanometers. The state is formed by the evolution of GBs into three-dimensional (3D) minimal-interface structures constrained by twin boundary networks. In this type of polycrystal, grain coarsening is effectively suppressed, even close to the melting point. The strength is also close to the theoretical value of Cu.

With a two-step plastic deformation process of surface mechanical grinding treatment followed by high-pressure torsion in liquid nitrogen, grains of polycrystalline Cu with a purity of 99.97 wt % were refined into the nanometer scale (see the supplementary materials). Bright-field transmission electron microscopy (TEM) images of the extremely fine grains in the as-prepared specimens showed a typical manifold structure (Fig. 1A), as is normally observed in the bicontinuous lipid-water phases (*8*). The specimen consisted of irregularly shaped aggregates or chains connected to each other that form continuous networks. By increasing the magnification, we observed that the aggregates are made up of several individual grains of a few nanometers in size. Among grains of similar contrast in an aggregate, we detected either twin relationships or coincidence site lattice (CSL) boundaries. For example, as indicated by the fast Fourier transform (FFT) patterns of an aggregate consisting of five grains (Fig. 1B), we observed two sets of Σ3 boundaries on $(1\bar{1}1)$ and $(\bar{1}\bar{1}1)$ planes, respectively, and a CSL boundary with a misorientation of 21° between G4 and G5.

TEM images obtained in high-resolution mode of the ultrathin foil specimens showed nanosized crystallites with various orientations in the as-prepared Cu (Fig. 1C). Grains were roughly equiaxed with a uniform size distribution, with most grains <10 nm in size. Selected area electron diffraction covering a large number of grains showed nearly continuous diffraction rings of face-centered cubic (fcc) Cu, an indication of seemingly random orientations of the polycrystallites. The tiny crystallites are connected with each other by atomically distinct boundaries. We did not detect amorphous phases or pores. The full crystallinity of the as-prepared Cu sample is in concert with x-ray diffraction patterns, where sharp diffraction peaks show up without a diffused amorphous peak.

Our procession electron diffraction measurements under TEM (Fig. 1D) revealed that a considerable fraction of GBs (~25%) are Σ3 twin boundaries with a misorientation of 60° in the sample. Other interfaces separating grains have randomly distributed misorientations, among which different types of low-Σ CSL boundaries were detected as well, constituting ~20%. This is in agreement with our other observations (Fig. 1, A and B). As in the magnified images (Fig. 1E), some boundaries are associated with GB structural units as previously observed, underlying the typical structures of relaxed GBs. All of these structural characteristics support the assumption that the GBs are in relaxed states, which is consistent with the decreased GB energy in the relaxed nanograined Cu below the critical grain size (*6*).

We characterized individual grains by properly tilting the specimens under high-resolution TEM so as to resolve their lattice images. We identified diverse geometries for plenty of individual grains. Rather than curved boundaries, many grains distinguish themselves by faceted boundary planes, such as (111) and (100) atomic planes. We show a typical grain of ~2 nm in diameter imaged along the [110] zone axis that is shaped by four faceted {111} and two {001} boundaries (Fig. 2A). Not all faceted boundaries are twin boundaries. Some grains are defect free, whereas others contain

[1]Shenyang National Laboratory for Materials Science, Institute of Metal Research, Chinese Academy of Sciences, Shenyang 110016, China. [2]School of Materials Science and Engineering, Shanghai Jiaotong University, Shanghai 200240, China.
*Corresponding author. Email: xyli@imr.ac.cn (X.Y.L.); lu@imr.ac.cn (K.L.)

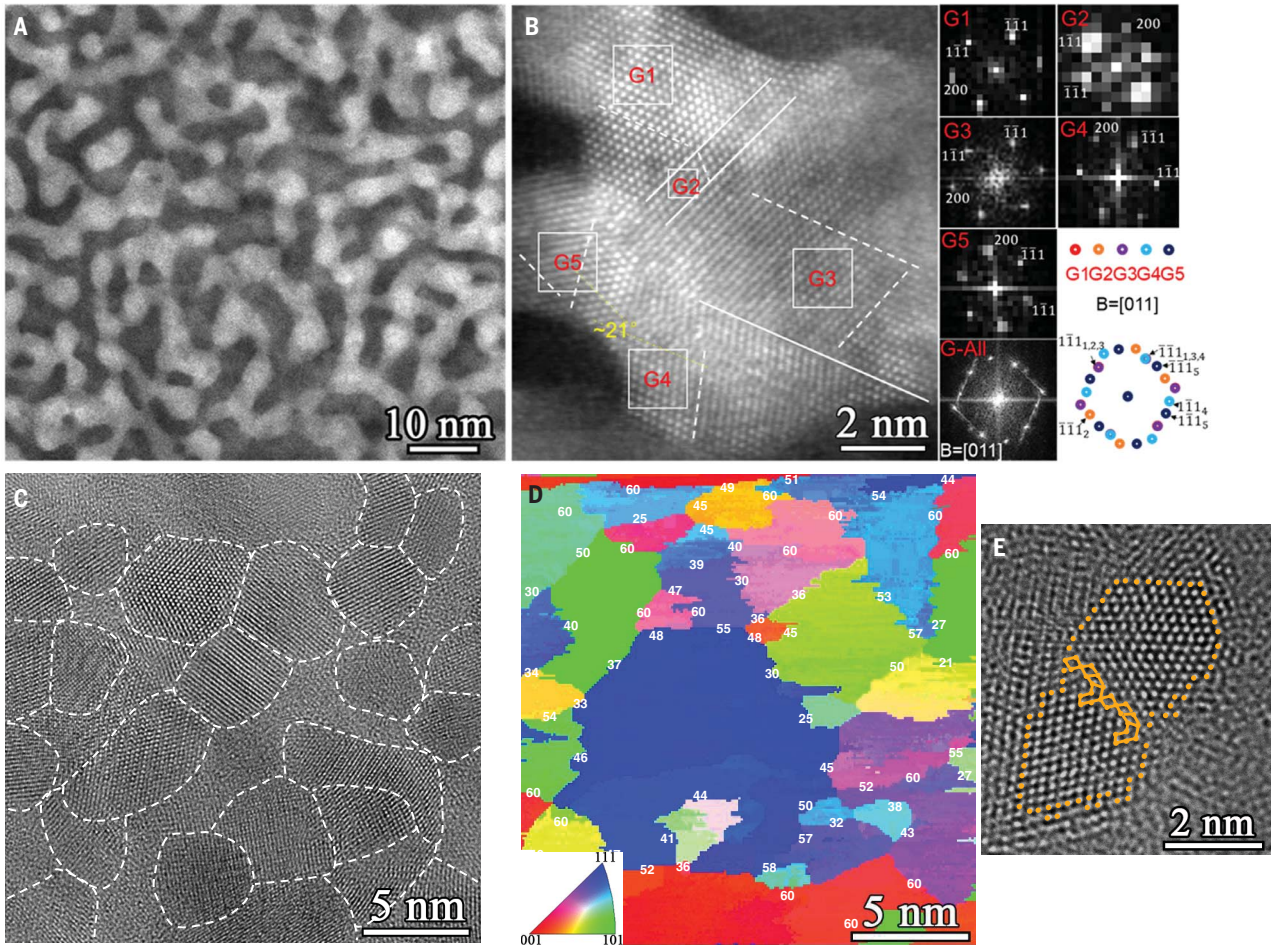


Fig. 1. Microstructures of the as-prepared Cu sample with extremely fine grains. (**A**) A typical bright-field TEM image. (**B**) (Left) A magnified image of a selected area in (A). Dashed lines represent {111} planes and solid lines show CTBs. (Right) Corresponding FFT images of grains (G1, G2, G3, G4, and G5) labeled in the left panel. G-All indicates all grains, with a schematic on the right. (**C**) A typical high-resolution TEM image. (**D**) A typical inverse pole figure (IPF) image acquired from a region in (C) from the precession electron diffraction analysis. Numbers indicate misorientation angles of GBs. (**E**) A typical structural-unit–type boundary as outlined between two tiny grains.

twins or stacking faults that cut through the grains. The twinned or faulted grains are also bounded by {111} and {001} atomic planes (Fig. 2, C and E). The {111} twin boundaries in the grains are fully coherent, with spacing of only a few atomic layers. Statistically, for grains <3 nm in size, 62% of faceted boundaries are found to be {111} and 33% are found to be {100}—a frequency ratio of about 2:1. For grains between 3 and 10 nm, the numbers are 52% {111} and 28% {100}, respectively. This feature disappears in grains >10 nm, which supports the idea that grain refining facilitates the formation of faceted {111} and {100} boundaries at the expense of other boundaries.

We also found that the shapes of those grains resemble that of the truncated octahedron (*9*). The atomic images of the defect-free grain (Fig. 2A) can be perfectly fitted with a part of a truncated octahedron projected along the [110] zone axis, except for several corners where atomic images are missing (circled in Fig. 2A). Other grains (Fig. 2, C and E) also match the twinned or faulted polyhedra very well. Lattice spots on the two corners of the grain in Fig. 2C (arrows) are blurred or absent in the TEM image. We identified a stacking fault zone between two truncated-octahedron–shaped grains (Fig. 2E), which is consistent with other observations (Fig. 1). Along the [110] projection of a truncated octahedron, four {111} and two {100} planes exist, with a frequency ratio of 2:1. This ratio is the same as what we observed in the experiments. The truncated-octahedral shape is apparently a favorable option for grains <10 nm.

We determined the thermal stability of the as-prepared Cu samples with an average grain size of 10 nm by isothermal annealing at various temperatures. We detected no apparent grain coarsening with increasing the annealing temperature up to 1348 K, which is only 9 K below the equilibrium melting point of Cu (1357 K) (Fig. 3A and fig. S3). Holding for 15 min at this temperature and cooling down, the extremely fine grains remained (Fig. 3, B and C). We still observed polyhedron-shaped grains with an average size of ~11 ± 2.3 nm with distinct boundaries—akin to what we observed before annealing. We detected more twins in the annealed grains (Fig. 3C), presumably because of the further dissociation of GBs during annealing at elevated temperatures. Elevating the temperature above 1357 K induced melting, at which point all nanograins disappeared. Our observations lead us to the conclusion that no substantial grain coarsening seems to occur before melting.

We prepared a separate sample for comparison with larger grains (average size of 25 nm) using the same process (see the supplementary materials) but with a smaller strain. We observed the onset of grain coarsening in this sample at ~985 K (Fig. 3A). The grain coarsening temperature is higher than that of 50-nm grains, which is a trend that runs contrary to the conventional belief that smaller grains coarsen at lower temperatures. Our observations support the idea that GB relaxations in polycrystals of

smaller grain size may operate more effectively to enhance the thermal stability (*6*). The Cu samples in this study are the most stable compared with the literature data for Cu (*10*–*19*), including nanograined, deformed coarse-grained, or the metastable amorphous Cu-based alloys with crystallization temperatures in the range of 0.33 to 0.36 T_m.

For the 10-nm grain–sized sample, we measured nanoindentation hardness to be as high as 4.7 ± 0.2 GPa (fig. S2). This value is equivalent to a yield strength of 1.57 GPa. This value is also higher than reported data on nanograined Cu specimens as well as amorphous Cu-based alloys (*20*–*26*). The ideal shear strength of Cu at 0 K is 2.16 GPa. The shear strength we estimated from the nanoindentation is nearly comparable to the ideal shear strength when temperature effects are considered (*27*). From previous measurements, the strength of different Cu samples follows the classical Hall-Petch relationship for grains >30 nm. For finer grains, a strength plateau or even softening occurs (*20*–*23*). By contrast, we did not find this trend in our hardness measurements. The strengthening we observed suggests that plastic deformation events, such as GB migration or sliding, must be effectively prohibited in the present Cu sample. We also found no trace of dynamic recrystallization.

The unusual stability we observed supports the belief that when grains are refined to the extreme scale, the polycrystalline structure seems to readily adopt one or more specific states. Our TEM observations, combined with the reconstructed atomic models, indicate that the truncated octahedron—either perfect or containing planar faults—is a prototype structure that represents a large portion of individual grains during the deformation-induced refinement process. Accordingly, potential structures and GB networks can be perceived on the basis of the well-known Kelvin conjecture—that is, a topological form of grains representing polycrystalline metals constructed with equal-sized grains of truncated-octahedral shape. This model provides a simple geometrical solution of idealized 3D polycrystals with nearly the minimum total area of GBs (*28*). Because a considerable fraction of GBs are Σ3-type twin boundaries in our as-prepared sample, the structure of a 3D space–filling twin boundary network must be considered as a basic precept as well (*29*).

We set up an atomistic model aimed at addressing the outstanding stability of the extremely fine Cu grains. Referring to the Kelvin model, which consists of two ideal truncated-octahedra of equal size with their centers (K1 and K2; Fig. 4A) in a body-centered cubic array, we constructed an extended Kelvin supercell with 16 truncated-octahedral–shaped grains of equal size. We can recognize

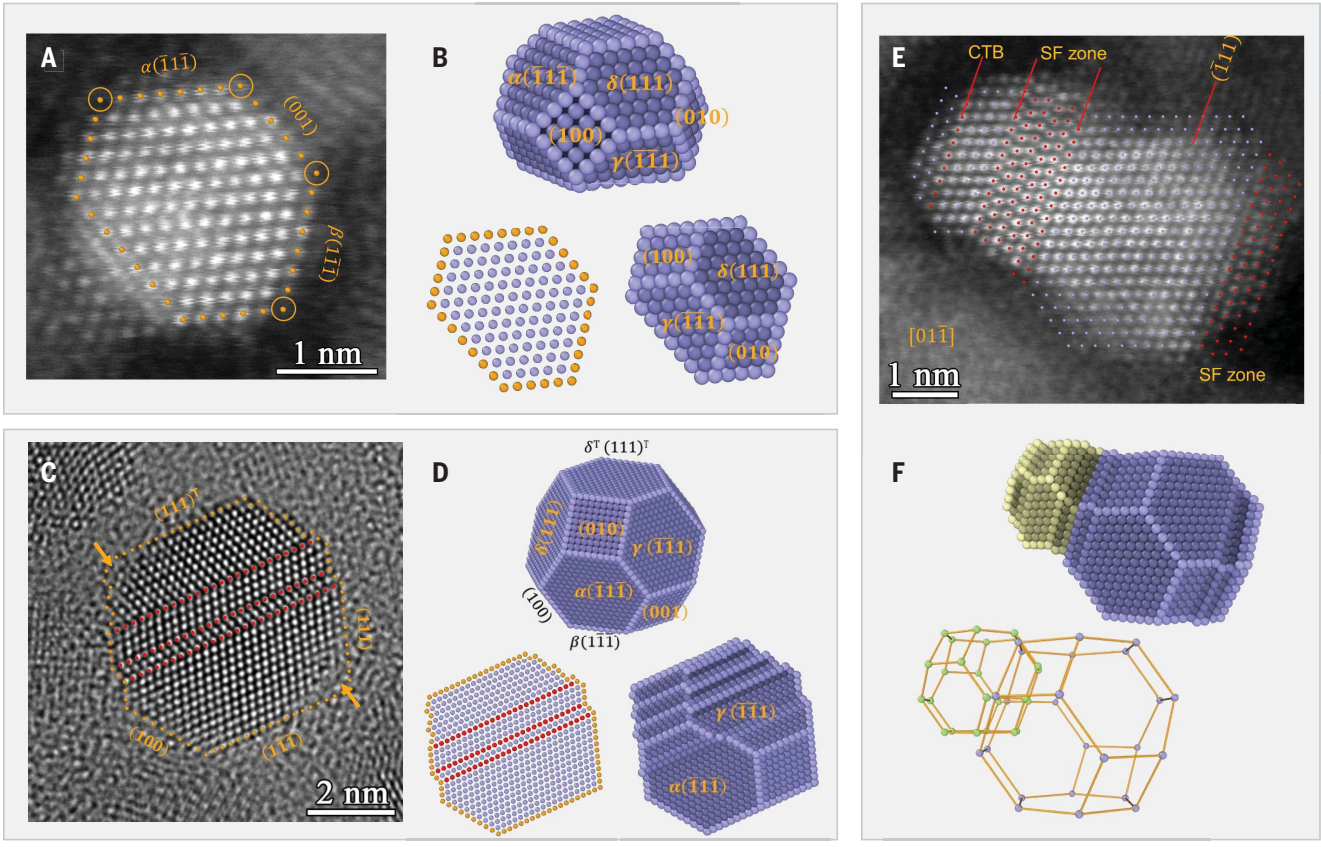


Fig. 2. High-resolution TEM images of individual grains with truncated-octahedral geometries. (**A**) A tiny grain of ~2 nm in size. (**B**) A part of an ideal truncated-octahedron with 1154 atoms (top), rotated by 49° along the [110] axis (lower right). The projected atomic positions on the (001) plane (lower left) are coincident with the TEM image in (A) (where only border atoms are shown in orange). Corner atoms in blurred contrast are circled in (A). (**C**) A grain containing twins. (**D**) An ideal truncated-octahedron of 11,817 atoms (top), rotated by 25.5° about the [01$\bar{1}$] axis after introducing twins (lower right). Projected atomic positions (bottom left) agree with the TEM image in (C) (where only border atoms in orange and twin boundary atoms in red are shown). Missing corners are indicated by orange arrows in (C). (**E**) Two grains containing stacking faults (SFs) and twins. (**F**) Two attached truncated-octahedral grains of different sizes with projected atomic positions agreeing with the TEM image in (E).

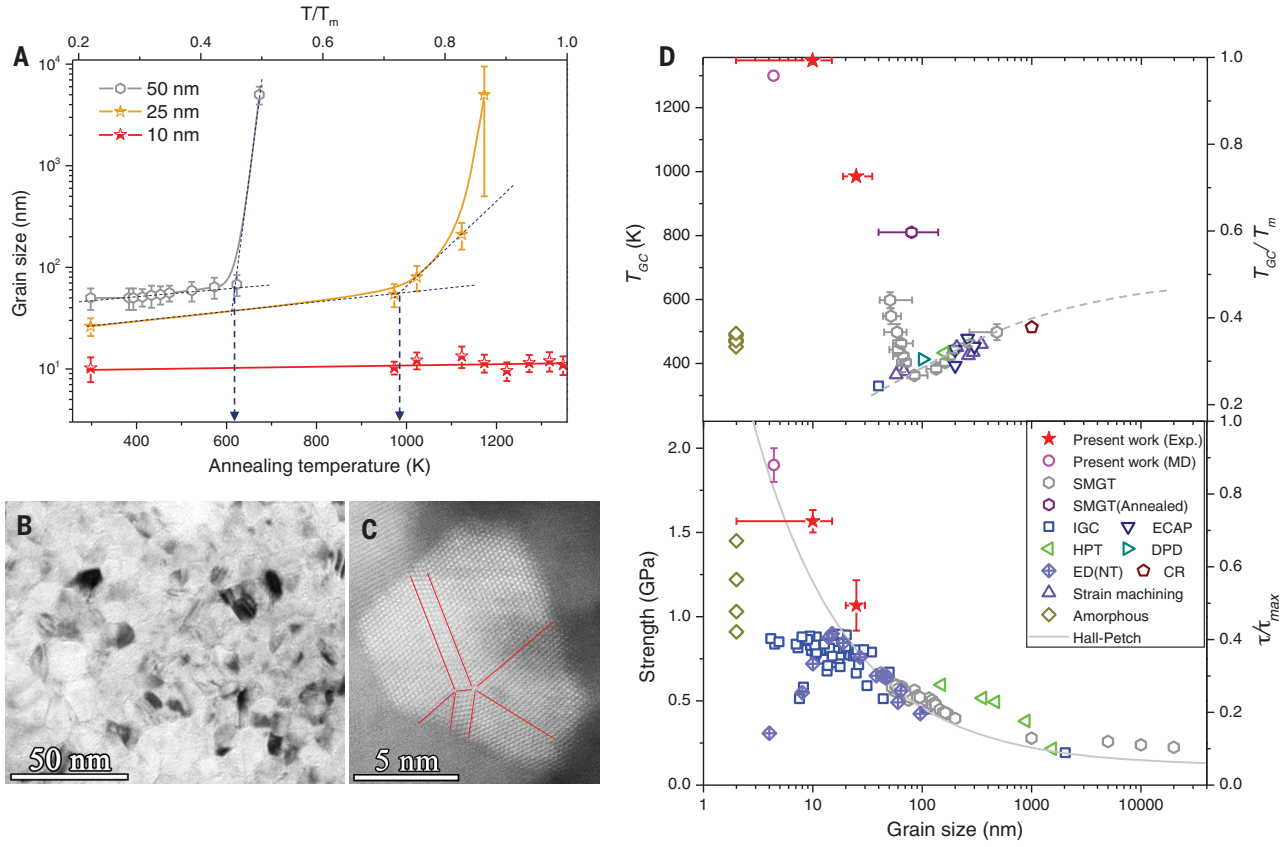


Fig. 3. Extremely high thermal stability and strength. (**A**) Grain size variations as a function of annealing temperature for three samples with initial average grain sizes of 50 nm, 25 nm, and 10 nm, respectively. Each point of grain size was averaged from >300 grains. (**B**) A TEM image of the sample with an initial grain size of 10 nm after annealing at 1348 K for 15 min. (**C**) A high-resolution TEM image of a grain in (B). Red lines indicate twin boundaries. (**D**) Grain coarsening temperatures (T_{GC}) and strength as a function of grain size in pure Cu. Literature data for Cu samples prepared through various processes in (*6*, *10*–*17*, *20*–*24*) are included. Data for amorphous Cu alloys are from (*18*, *19*, *25*, *26*). T_m, melting point of Cu; τ_{max}, ideal shear strength of Cu (*27*). Each grain coarsening temperature was obtained from three independent experiments, and each strength datum was obtained from 10 independent experiments. Exp., experimental; SMGT, surface mechanical grinding treatment; IGC, inert gas condensation; ECAP, equal channel angular pressing; HPT, high-pressure torsion; DPD, dynamic plastic deformation; ED(NT), electrodeposition (nanotwin); CR, cold rolling.

the fundamental characteristics of GBs, and twin boundary networks can be recognized. By taking polyhedron O1 as the center cell, we can specify the lattice orientation of its eight nearest neighboring grains such that each {111} face of O1 can form a Σ3 coherent twin boundary (CTB) with them—i.e., A1, B1, C1, D1 (upper) and A2, B2, C2, D2 (lower). The lattice orientations of the six second-nearest neighboring grains to O1 (I1, E1, F1 and I2, E2, F2) are not strongly limited. We chose them such that in each <001> orientation, a Σ9-type tilt GB is formed with O1 (or O2) (fig. S4 and table S2). Our result confirms the theoretical prediction (*29*) that a 3D space–filling CTB network can be constructed by a specific arrangement of truncated octahedra in fcc solids.

By choosing an extended Kelvin polycrystal with an initial grain size of 3.27 nm as a starting structure for simplicity, we carried out MD simulations to relax the sample by heating it up to different target temperatures at a rate of 8 K/ns. We based the interatomic potential on the semiempirical model developed with the embedded atom methods (EAM) (*30*). For comparison, we also conducted MD tests on a number of samples representing cases for flat and curved GBs, in which both grain size and orientations are treated as free variables (fig. S6).

During MD relaxation and subsequent heating, GBs in the extended Kelvin polycrystal transform into different structures through extensively GB reaction and migration events. With increasing temperature, GB reactions—such as dissociation and coalescence of GBs associated with some Σ9-type GBs—are triggered at triple lines or junctions of GBs, as expected experimentally (*31*) and theoretically (*29*). Migration of twist GBs becomes substantial at mediate temperatures (640 K), and it is close to that for curved GBs observed in separate simulations (fig. S6). The preexisting CTB network is stable against thermal fluctuations. After heating to 760 K, although some grains shrank and finally disappeared because of GB migration, the entire GB network failed to collapse. The GBs merged and developed into a different form that topologically resembles the Schwarz D-surface (*32*, *33*). As a triply periodic minimal surface, we can identify the shape of the Schwarz D-interface in our calculations (Fig. 4B and fig. S5).

This transformation is thermodynamically driven by minimizing the interface area, as indicated by the MD results. Although the Kelvin polycrystal provides a reasonable solution for an efficient, space-filling arrangement of grains, the total boundary area was further minimized by forming the Schwarz D-interface structure (*34*). Our observation indicates that the polycrystalline structure with Schwarz D-interfaces is more stable than the Kelvin polycrystals, but this does not necessarily mean that such a transformation occurs at a distinct point during the refinement of coarse grains into the nanograins. Our MD simulations also confirmed that the Schwarz D-structure can be obtained by imposing strain at low temperatures. The

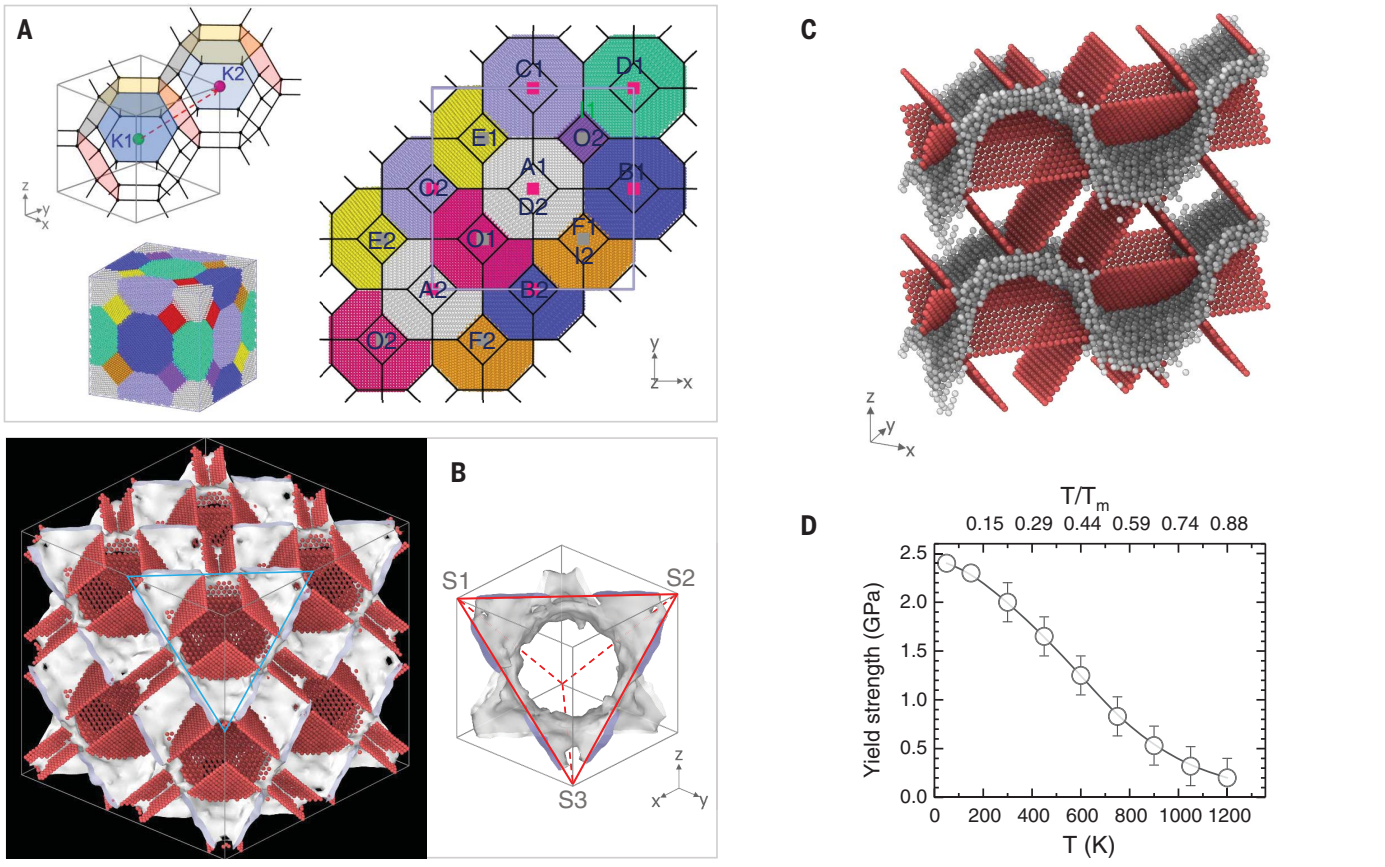


Fig. 4. Atomistic model and MD simulations of Schwarz crystals. (**A**) The original Kelvin model of two ideal truncated-octahedra of equal volume (K1 and K2) in 1 by 1 packing (top left). A polycrystal of 16 grains (right) was constructed using a 4 by 4 packing Kelvin model (initial grain size, 6.6 nm). A space-filling 3D CTB network was constructed with a specified lattice orientation for individual grains (see Supplementary Materials). (**B**) (Left) MD-obtained twin-bounded polycrystalline structure at 0 K, demonstrated by 2 by 2 by 2 supercells where atoms in fcc lattice sites are removed. (Right) GBs resembling the Schwarz D-interface in a 1 by 1 by 1 supercell. (**C**) A section-view of the Schwarz crystal showing Schwarz D-GBs constrained by CTB networks. (**D**) The MD-obtained yield stress as a function of temperature. Error bars quantify uncertainty caused by rate effects and thermal fluctuations.

adoption of the minimal-interface structure constitutes an energetically more-favorable GB relaxation process under thermal and mechanical stimuli.

The Schwarz D-interface is constrained by high-density CTBs on both sides (Fig. 4C)—a structure meant to suppress GB migration. With the minimal-interface area and the zero-mean curvature (another characteristic of minimal interface structures), the driving force to migrate the GBs tends to vanish, which is proportional to the GB curvature. CTBs are intrinsically stable at elevated temperatures. The stable CTB networks and Schwarz D-interface interlock with each other, preventing GBs from developing substantial curvatures to escape and CTBs from migrating. Such a pronounced stability cannot be achieved with the Schwarz D-interface only, nor with only the nanotwinned structures.

The obtained Schwarz D-structure stays stable at elevated temperatures. Rather than grain coarsening, roughening of GBs occurs as the melting point is approached, at which point the liquid phase is nucleated heterogeneously at 1321 K (fig. S6). This suggests that the upper thermal stability is limited kinetically by GB melting. This notable stability agrees fully with our experimental results (Fig. 3). Clearly, preventing the migration of GBs through a minimal interface constrained by a 3D CTB network constitutes the key mechanism to achieve the pronounced stability for polycrystals at such extreme sizes.

We conducted uniaxial tensile-loading tests on the CTB-constrained Schwarz D-structure at various temperatures and strain rates (10^5 to 10^7 s^{-1}) (fig. S7). GB activities such as sliding or migration can be suppressed entirely. The primary mode of deformation is twinning. For example, a twinning partial dislocation is generated from one GB site and absorbed at the opposite GB, which results in migration of the twin boundary plane. Generating a stacking fault to a glide plane parallel to the CTB is also possible during the plastic deformation. The critical stress corresponding to incipient twinning is temperature-dependent (Fig. 4D). The yield strength is 1.8 to 2 GPa at 300 K, which is higher than our experimentally measured data, possibly owing to the grain size distribution in real samples. The strength at 10 K is comparable to the ideal pure shear strength obtained through first-principle calculations (*27*).

Evidence based on our experiments and MD simulations confirms that a pronounced stability is achievable if the minimal-interface structure constrained by a CTB network is adopted in polycrystalline Cu with nanosized grains. The structure, which we refer to as a Schwarz crystal, is a different kind of metastable state for polycrystalline solids. This state differs fundamentally from the amorphous solid states. Because the metastable structure is formed when twinning—either thermally induced or deformation induced—and is effective in relaxing GB structures, the appearance of the Schwarz crystal in different metals and alloys is expected with the activation of twinning mechanisms at

the nanoscale. The pure Cu Schwarz crystal, which contains a very high density of interfaces, exhibits a thermal stability as high as that of a single crystal and much higher than that of amorphous solids. Therefore, the structure provides emerging opportunities for exploring physical and chemical phenomena of metals, particularly with respect to transport dynamics of atoms and electrons at interfaces and interactions of various defects at high temperatures. As a benchmark structure capable of suppressing grain coarsening under both thermal fluctuations and external forces, the Schwarz crystal allows for elevated stability and strength that goes with refining grains into the extremely fine scale. This overcomes some of the difficulties present in traditional strategies for materials development. In principle, the Schwarz crystal should be accessible in other materials and may provide a different direction for developing strong and stable materials for high-temperature applications.

REFERENCES AND NOTES

1. R. Zallen, *The Physics of Amorphous Solids* (Wiley, 1983).
2. P. Haasen, B. L. Mordike, *Physical Metallurgy* (Cambridge Univ. Press, 1996.)
3. K. Lu, *Nat. Rev. Mater.* **1**, 16019 (2016).
4. P. Färber, E. Cadel, A. Menand, G. Schmitz, R. Kirchheim, *Acta Mater.* **48**, 789–796 (2000).
5. C. A. Schuh, T. G. Nieh, H. Iwasaki, *Acta Mater.* **51**, 431–443 (2003).
6. X. Zhou, X. Y. Li, K. Lu, *Science* **360**, 526–530 (2018).
7. X. Zhou, X. Y. Li, K. Lu, *Phys. Rev. Lett.* **122**, 126101 (2019).
8. W. Longley, T. J. McIntosh, *Nature* **303**, 612–614 (1983).
9. C. Chieh, *Acta Cryst.* **A35**, 946–952 (1979).
10. C. Saldana, A. H. King, S. Chandrasekar, *Acta Mater.* **60**, 4107–4116 (2012).
11. X. Y. Li, X. Zhou, K. Lu, *Sci. Adv.* **6**, eaaz8003 (2020).
12. A. Kumpmann, B. Günther, H. D. Kunze, *Mater. Sci. Eng. A* **168**, 165–169 (1993).
13. N. Lugo, N. Llorca, J. J. Suñol, J. M. Cabrera, *J. Mater. Sci.* **45**, 2264–2273 (2010).
14. C. F. Gu, C. H. J. Davies, *Mater. Sci. Eng. A* **527**, 1791–1799 (2010).
15. Y. Zhang, N. R. Tao, K. Lu, *Acta Mater.* **56**, 2429–2440 (2008).
16. P. Jenei, J. Gubicza, E. Y. Yoon, H. S. Kim, J. L. Lábár, *Compos. A: Appl. Sci. Manuf.* **51**, 71–79 (2013).
17. Z. N. Mao *et al*., *Mater. Sci. Eng. A* **674**, 186–192 (2016).
18. B. W. Zhang, X. L. Shu, S. Y. Zhu, S. Z. Liao, *J. Mater. Proc. Tech.* **91**, 90–94 (1999).
19. H. Zhang, Z. Tan, B. Zhang, X. Shu, J. Yu, *J. Mater. Sci. Lett.* **10**, 45–46 (1991).
20. L. Lu, X. Chen, X. Huang, K. Lu, *Science* **323**, 607–610 (2009).
21. A. H. Chokshi, A. Rosen, J. Karch, H. Gleiter, *Scr. Metall.* **23**, 1679–1683 (1989).
22. G. E. Fougere, J. R. Weertman, R. W. Siegel, *Nanostruct. Mater.* **3**, 379–384 (1993).
23. P. G. Sanders, J. A. Eastman, J. R. Weertman, *Acta Mater.* **45**, 4019–4025 (1997).
24. H. G. Jiang, Y. T. Zhu, D. P. Butt, I. V. Alexandrov, T. C. Lowe, *Mater. Sci. Eng. A* **290**, 128–138 (2000).
25. W. H. Wang, *J. Appl. Phys.* **99**, 093506 (2006).
26. Y. P. Deng *et al*., *Intermetallics* **15**, 1208–1216 (2007).
27. S. Ogata, J. Li, S. Yip, *Science* **298**, 807–811 (2002).
28. F. N. Rhines, K. R. Craig, R. T. DeHoff, *Metall. Trans.* **5**, 413–425 (1974).
29. V. Y. Gertsman, B. W. Reed, *Z. Metallk.* **96**, 1106–1111 (2005).
30. V. Borovikov, M. I. Mendelev, A. H. King, R. LeSar, *Model. Simul. Mater. Sci. Eng.* **23**, 055003 (2015).
31. T. Yu, N. Hansen, X. Huang, *Proc. R. Soc. A.* **467**, 3039–3065 (2011).
32. H. A. Schwarz, *Gesammelte Mathematische Abhandlungen*, vols. 1 and 2 (Springer, 1890).
33. A. L. Mackay, *Nature* **314**, 604–606 (1985).
34. L. E. Scriven, *Nature* **263**, 123–125 (1976).

ACKNOWLEDGMENTS

We thank C. Y. Wu for help with sample preparation. **Funding:** This work is funded by the Ministry of Science and Technology of China (grant nos. 2017YFA0204401 and 2017YFA0700700) and the Chinese Academy of Sciences. **Author contributions:** X.Y.L. and K.L. developed the concept, designed the experiments, and supervised the projects. X.Z. and X.Y.L. prepared the samples and characterized structures. X.Z. measured properties. Z.H.J. performed modeling and simulations. All authors discussed and analyzed the results. X.Y.L., Z.H.J., and K.L. wrote the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in the manuscript or the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/831/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S3
References (*35*–*44*)

1 August 2020; accepted 8 October 2020
10.1126/science.abe1267

COSMOCHEMISTRY

Astronomical context of Solar System formation from molybdenum isotopes in meteorite inclusions

Gregory A. Brennecka[1,2]*, Christoph Burkhardt[2], Gerrit Budde[2,3], Thomas S. Kruijer[1,4], Francis Nimmo[5], Thorsten Kleine[2]

Calcium-aluminum–rich inclusions (CAIs) in meteorites are the first solids to have formed in the Solar System, defining the epoch of its birth on an absolute time scale. This provides a link between astronomical observations of star formation and cosmochemical studies of Solar System formation. We show that the distinct molybdenum isotopic compositions of CAIs cover almost the entire compositional range of material that formed in the protoplanetary disk. We propose that CAIs formed while the Sun was in transition from the protostellar to pre–main sequence (T Tauri) phase of star formation, placing Solar System formation within an astronomical context. Our results imply that the bulk of the material that formed the Sun and Solar System accreted within the CAI-forming epoch, which lasted less than 200,000 years.

Calcium-aluminum–rich inclusions (CAIs) are the oldest dated samples of the Solar System and provide a direct record of its formation. These micrometer- to centimeter-sized inclusions in meteorites formed in a high-temperature environment (>1300 K), probably near the young Sun (*1*), and were subsequently transported outward to the region where carbonaceous chondrite meteorites (and their parent bodies) formed, in which they are found today (*2*–*4*). Although some peculiar CAI-like objects, with intrinsic percent-level isotope differences from other Solar System materials, may have formed at a slightly earlier stage [e.g., (*5*)], the majority of CAIs formed 4567.2 ± 0.2 million years (Myr) ago, over a period of ~40,000 to 200,000 years (*6*–*8*). Astronomical observations of young stellar objects (YSOs) show that solar-mass stars accrete by infall from their parental molecular cloud cores before transitioning into pre–main sequence stars surrounded by a disk of gas and dust (referred to as classes 0 to II in the star formation literature) within ~1 Myr (*9*). This is longer than CAIs took to form, raising the question of which astronomical phase in the Solar System's formation is recorded by the formation of CAIs. Previous work modeling the formation of the Solar System and observations of other stellar systems have associated the formation of CAIs either with the period of collapse and accretion of a protostar (class 0 phase) (*1*, *6*) or with the pre–main sequence, T Tauri phase (class II phase) of the proto-Sun, which would have occurred up to ~1 Myr later (*10*, *11*). Distinguishing between these interpretations would provide a link between astronomical observations of YSOs and the physical samples defining the age of the Solar System.

We investigate this link using the isotopic compositions of CAIs. Previous work has shown that CAIs are isotopically distinct from subsequently formed solids in the solar protoplanetary disk (*12*, *13*), but any "genetic" connections between CAIs and later-formed solids are unclear. Compared with bulk meteoritic materials, CAIs are ^{16}O-rich and more similar to the Sun than later-formed objects (*14*). However, because O-isotope variations reflect the complex interplay of mixing and fractionation processes within the protoplanetary disk, and perhaps the parent molecular cloud (*14*), O isotopes alone are of limited use for linking CAI formation to stellar accretion and disk formation. Another type of isotope variation, known as nucleosynthetic variation, is due solely to inherent isotopic differences in the starting material and can be used to identify temporal and spatial differences in the provenance of Solar System materials. These isotope variations arise from the heterogeneous distribution of presolar material within the disk and show that the protoplanetary disk can be divided into noncarbonaceous (NC) and carbonaceous (CC) reservoirs, which have been suggested to represent the inner and outer Solar System, respectively (*15*–*17*). This NC-CC isotopic dichotomy has been demonstrated in several elements, including Mo, which can distinguish between isotope variations arising from the heterogeneous distribution

[1]Lawrence Livermore National Laboratory, Livermore, CA, USA. [2]Institut für Planetologie, University of Münster, Münster, Germany. [3]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA. [4]Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Berlin, Germany. [5]Department of Earth & Planetary Sciences, University of California Santa Cruz, Santa Cruz, CA, USA.
*Corresponding author. Email: brennecka2@llnl.gov

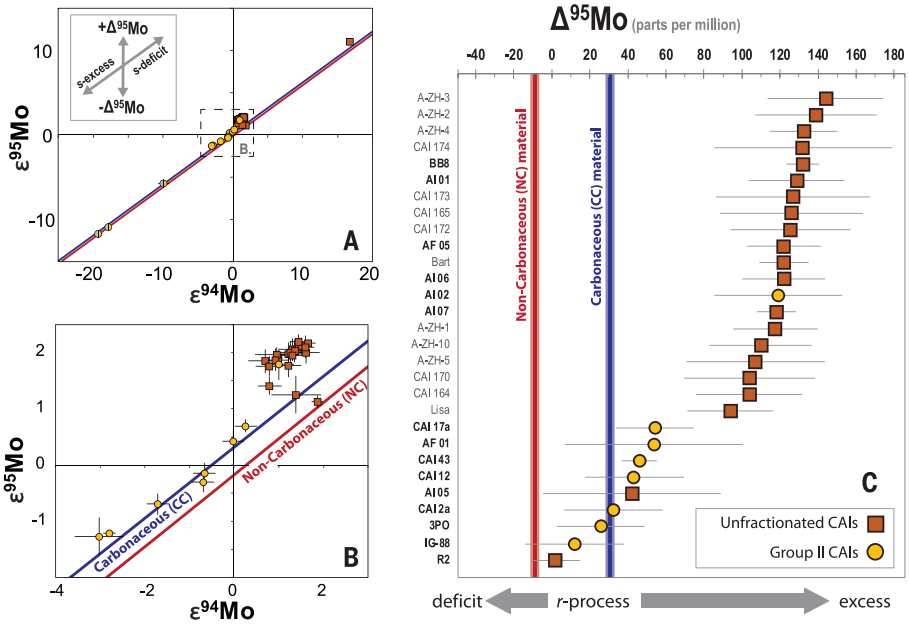


Fig. 1. Molybdenum isotopic data for all CAIs in our sample set. (**A**) ε^{94}Mo versus ε^{95}Mo [where ε notation represents parts per 10,000 deviation from a terrestrial standard, (*20*)]. Data plotted are from this work and other studies (*12*, *18*, *25*). Samples with half-gray symbols have been excluded from calculations of Δ^{95}Mo (*20*). The CC (blue) and NC (red) correlations (*21*) are shown for reference. (**B**) Magnified view of the region within the dashed box in (A). (**C**) CAI isotopic variability due to just the r-process, presented using Δ^{95}Mo notation (see text for details). The left axis lists the CAIs from literature sources (*12*, *18*, *25*) in plain text and those from this study in bold. We have excluded CAIs from the literature that do not have comparable methodologies and CAIs with terrestrial contamination of their Mo isotopic composition. Error bars show 95% confidence intervals. The CC (blue) and NC (red) correlations (*21*) are shown for reference. In (A) to (C), unfractionated CAIs are plotted with orange squares and group II CAIs with yellow circles.

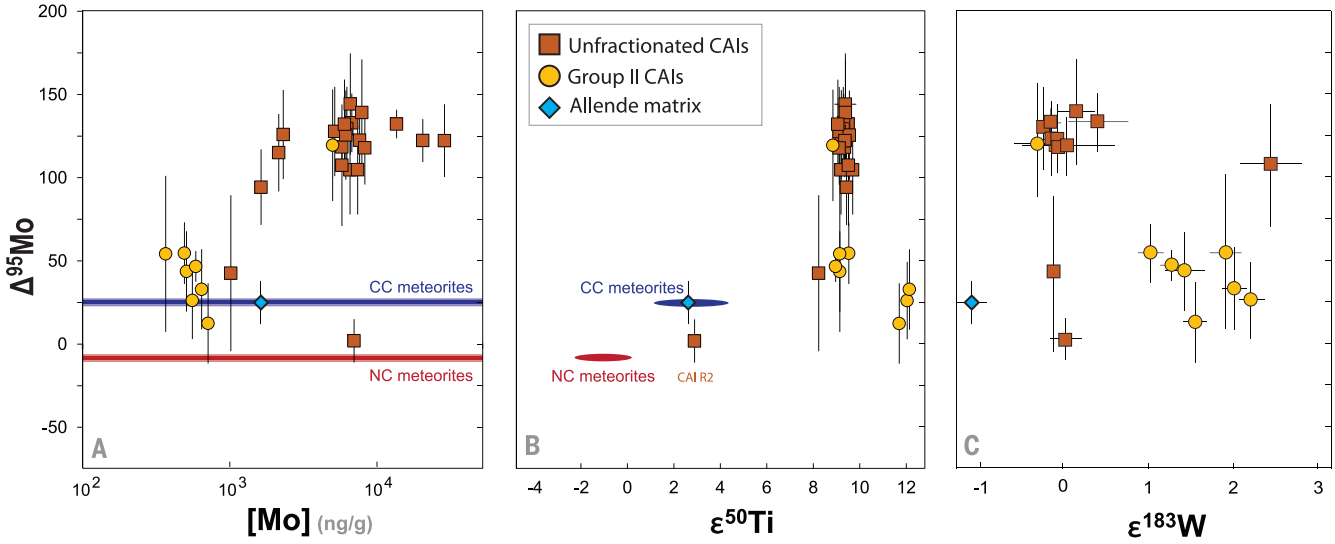


Fig. 2. Isotopic and concentration correlations in CAIs. (**A**) Δ^{95}Mo plotted against concentration of Mo in the CAIs. In general, samples with a lower concentration of Mo also have a lower Δ^{95}Mo, suggesting at least some disk processing or CAI formation effect on the r-process signature of the CAI. (**B**) Δ^{95}Mo plotted against ε^{50}Ti in CAIs. See text for discussion of these isotopic signatures in CAIs. (**C**) Δ^{95}Mo plotted against ε^{183}W. The Δ^{95}Mo and ε^{183}W values do not correlate with one another or show a mixing relationship with the meteorite matrix. This indicates that parent body alteration does not control these signatures. Bulk NC and CC meteorite data for Mo are from (*21*), Ti data are from (*31*, *32*), and the matrix of the CC meteorite Allende (blue diamonds) is from (*16*, *28*). Other plotting symbols are as in Fig. 1.

of matter produced in the slow (s-process) and rapid (r-process) neutron capture processes of stellar nucleosynthesis (*15*, *16*, *18*, *19*). Whereas there are large s-process variations among materials within each group, all CC materials are characterized by an approximately constant r-process excess over NC materials. Deviations from pure s-process variation are quantified using the Δ^{95}Mo notation [$\Delta^{95}\text{Mo} \equiv (\varepsilon^{95}\text{Mo} - 0.596 \times \varepsilon^{94}\text{Mo}) \times 100$, where ε represents parts per 10,000 deviation from a terrestrial standard (*20*)], which isolates variation in the amount of r-process Mo from variation in the amount of s-process Mo. Previous studies have shown that the Δ^{95}Mo values for the NC reservoir, thought to represent the inner Solar System ($\Delta^{95}\text{Mo} = -9 \pm 2$), and CC reservoir, thought to represent the outer Solar System ($\Delta^{95}\text{Mo} = 26 \pm 2$), are distinct from one another (*21*). The NC-CC isotopic difference was established within the first ~1 Myr of Solar System history (*15*), so it likely reflects the incomplete mixing of isotopically variable material infalling from the Sun's protostellar envelope onto the disk at different times (*22*–*24*). Molybdenum isotopes, therefore, provide a tool for linking the formation of CAIs to the early processes of infall and disk building and, hence, to a specific stage of star formation.

We measured the Mo isotopic and trace element compositions for a variety of CAIs taken from the carbonaceous chondrite (CV3 group) meteorites Allende (13 CAIs), NWA 6870 (CAI AI01), and NWA 6717 (CAI AI02). We combine these measurements with similar data from three previous studies (*12*, *18*, *25*), which are summarized in table S1. CAIs are classified by their condensation histories, as inferred from their rare earth element (REE) systematics. Hereafter, we refer to samples with unfractionated REE patterns as "unfractionated CAIs," as distinct from "group II CAIs," which have fractionated REE patterns, indicating the loss of an ultrarefractory component (*26*). In addition to Mo, isotopic compositions of the CAIs were determined for several other elements [Ti, Fe, Ni, Sr, Ba, Nd, Sm, Er, Yb, Hf, and W; (*20*)], all of which are isotopically anomalous in CAIs. The 15 samples newly measured in this study are isotopically consistent with previously published studies of CAIs for these elements and exhibit larger anomalies than bulk meteorites; however, we concentrate on their Mo isotopes (Fig. 1). In general, unfractionated CAIs exhibit less variation in s-process Mo signatures (i.e., ε^{i}Mo values; Fig. 1, A and B) and mostly have higher and less variable Δ^{95}Mo values (Fig. 1C). By contrast, group II CAIs tend to have more variable s-process Mo signatures (Fig. 1, A and B) and lower Δ^{95}Mo values (Fig. 1C). However, a similar overall range in r-process material (>100 parts per million; Fig. 1C) is present in both types regardless of how much s-process variability exists, showing that variations in s-process and r-process Mo nuclides in CAIs are not directly coupled.

The CAIs exhibit different Δ^{95}Mo values, only some of which overlap with the values observed for bulk CC and NC meteorites (Fig. 1C). Unlike the absolute Mo isotope anomalies (i.e., ε^{i}Mo; Fig. 1, A and B) and the isotope anomalies for other elements (Fig. 2, B and C, and tables S2 and S4), the Δ^{95}Mo values provide a link between CAIs and the solid material of the disk as sampled by bulk meteorites. That is, the composition of CAIs varies from larger r-process excesses (more positive Δ^{95}Mo) toward more NC-like compositions.

We interpret the distinct Δ^{95}Mo values for these CAIs by relating them to models of how the Δ^{95}Mo values of the NC and CC reservoirs may have been established. These models demonstrate that during infall of the Solar System's parental molecular cloud, the isotopic composition of the infalling material changed from high Δ^{95}Mo values in the early stages to lower, NC-like Δ^{95}Mo values toward the final stage; and that as a result of viscous spreading of the initial disk, the bulk of the mass of the inner protoplanetary disk comes from the later infall, whereas the outer disk contains a higher proportion of early infalling material (*4*, *22*–*24*). The rapid outward transport of material during the initial disk-building process provides a mechanism for the transport of CAIs from near the Sun to the outer Solar System (*2*–*4*). Within this framework, the range of Δ^{95}Mo in CAIs extending from values characteristic of early infall (higher than the CC reservoir) to that of bulk meteorites may have three potential origins: (i) exchange of Mo with the host meteorite during alteration on the parent body; (ii) changes in the isotopic composition of infall, and, hence,

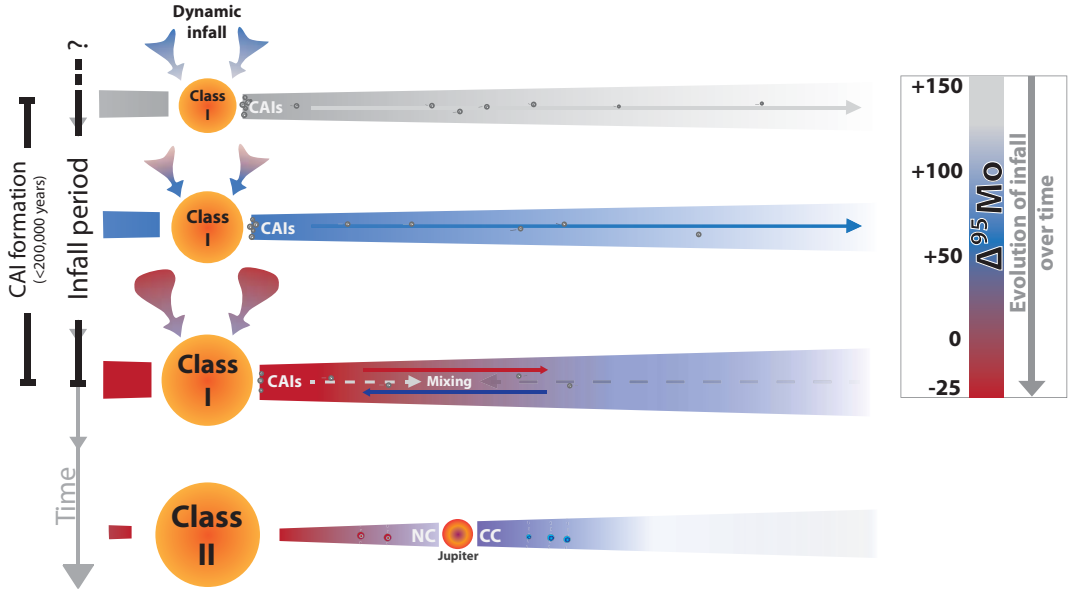


Fig. 3. Schematic diagram of the early evolution of the Solar System. The diagram is modeled after (*22*) but modified to reflect our interpretation of the CAI data and associated stages of stellar evolution. Heterogeneous infall from the Sun's parent molecular cloud evolves from a higher to lower amount of r-process material (i.e., higher to lower Δ^{95}Mo, indicated by the color bar). The isotopic compositions of CAIs record this change in composition and subsequent processing within the disk. The formation of CAIs occurs during the transition of the Sun from a class I to class II YSO in less than 200,000 years.

the gas from which CAIs condensed; or (iii) processing of CAIs within the disk before incorporation into their host meteorite.

Group II CAIs have substantially lower concentrations of Mo than unfractionated CAIs and so are more susceptible to alteration in their Mo isotopic composition than unfractionated CAIs. This may explain why group II samples tend to have lower (more bulk meteorite-like) Δ^{95}Mo values, although two unfractionated CAIs (AI05 and R2) also show such isotopic compositions. Nevertheless, because the Δ^{95}Mo values of many group II CAIs are similar to those of the matrix in Allende (*16*), their Δ^{95}Mo values may potentially reflect exchange with the host meteorite on the parent body. However, W, which is thought to behave similarly to Mo during parent-body alteration (*27*), is also strongly depleted in group II CAIs, yet all group II CAIs with low Δ^{95}Mo exhibit positive ε^{183}W signatures, unlike the negative ε^{183}W of the matrix of CV3 chondrites (*28*). This indicates that parent-body alteration is not the source of the signatures. As such, the combined Mo-W isotopic signatures of group II CAIs are difficult to account for solely by parent-body processes. This is also true for the large range of ε^{i}Mo values in group II CAIs (Fig. 1, A and B) and for the isotopic variation in unfractionated CAIs. Thus, although parent-body processes may have modified Mo isotope signatures in some group II CAIs, they cannot account for most of the isotopic variations that we see (see supplementary text). Consequently, the range of Δ^{95}Mo in CAIs was likely acquired before chondrite parent body accretion, either during CAI condensation or subsequent processing of CAIs within the disk.

If the range of Δ^{95}Mo in CAIs directly recorded the change of the isotopic composition of infall and, hence, of the gas from which CAIs condensed, CAIs should also directly record changes in other isotope systems that distinguish between NC and CC compositions, such as Ti. Bulk NC meteorites are characterized by negative ε^{50}Ti anomalies down to about −2, whereas the majority of CAIs have ε^{50}Ti excesses around +9 [e.g., (*29*)]. Thus, if evolving infall from the molecular cloud was the sole cause of isotopic anomalies in CAIs, then Δ^{95}Mo should coevolve with ε^{50}Ti in CAIs, assuming that Mo and Ti had similar carriers. However, with one exception—CAI R2, which exhibits both low Δ^{95}Mo and low ε^{50}Ti—no such clear coevolution is observed (Fig. 2B). A change of infalling matter toward NC-like compositions, which is characterized by ε^{183}W ~ 0, also does not provide a viable explanation for the positive ε^{183}W values of group II CAIs. Together, these observations suggest that dynamic infall from the molecular cloud (Fig. 3) cannot be the sole cause of the observed isotopic signatures in CAIs. We propose that even though dynamic infall initially generated the range of Δ^{95}Mo variations within the disk, additional processing within the protoplanetary disk established the measured isotopic signatures of CAIs. This processing evidently did not affect ε^{50}Ti, which shows little variation among the CAIs we studied but resulted in substantial and uncorrelated anomalies of ε^{i}Mo and ε^{183}W. The reasons for this disparate behavior during processing are unclear, but they may include different thermal properties of the carriers involved, size or density sorting of these carriers, and chemical fractionations during CAI formation, which made Mo and W more susceptible to isotopic modifications by processes within the disk.

The range of Δ^{95}Mo values in CAIs covers almost the entire compositional range observed among bulk meteorites, suggesting that CAIs acquired their isotopic signatures when the Sun's protoplanetary disk had largely formed. This links CAI formation to the transition of the Sun from a class I to a class II protostar, which is consistent with evidence of the short-lived isotope ^{10}Be in CAIs (*11*) and noble gas signatures from hibonite inclusions (*30*)—refractory inclusions thought to have formed before the more common CAIs—indicating that these refractory objects formed when the Sun was a T Tauri star.

Because the vast majority of CAIs are thought to have formed within ~40,000 to 200,000 years (*6*–*8*), the recorded isotopic character of CAIs also represents an amount of time in which that isotopic character was established. Interpreting the isotopic compositions of CAIs in the framework of disk evolution models, and with the knowledge that CAIs formed close to the young Sun in a known time window, we constrain the primary accretion phase (i.e., classes I and II) of stars like the Sun to less than 200,000 years, consistent with astronomical observations of YSOs (*9*).

REFERENCES AND NOTES

1. J. A. Wood, *Geochim. Cosmochim. Acta* **68**, 4007–4021 (2004).
2. E. Jacquet, S. Fromang, M. Gounelle, *Astron. Astrophys.* **526**, L8 (2011).
3. F. C. Pignatale, S. Charnoz, M. Chaussidon, E. Jacquet, *Astrophys. J. Lett.* **867**, 23 (2018).
4. L. Yang, F. J. Ciesla, *Met. Planet. Sci* **47**, 99–119 (2012).
5. L. Kööp *et al*., *Geochim. Cosmochim. Acta* **189**, 70–95 (2016).
6. J. N. Connelly *et al*., *Science* **338**, 651–655 (2012).
7. N. Kawasaki *et al*., *Earth Planet. Sci. Lett.* **511**, 25–35 (2019).
8. G. J. MacPherson, N. T. Kita, T. Ushikubo, E. S. Bullock, A. M. Davis, *Earth Planet. Sci. Lett.* **331-332**, 43–54 (2012).
9. L. Cieza *et al*., *Astrophys. J.* **667**, 308–328 (2007).
10. F. J. Ciesla, *Icarus* **208**, 455–467 (2010).
11. N. Dauphas, M. Chaussidon, *Annu. Rev. Earth Planet. Sci.* **39**, 351–386 (2011).
12. G. A. Brennecka, L. E. Borg, M. Wadhwa, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17241–17246 (2013).

13. N. Dauphas, E. A. Schauble, *Annu. Rev. Earth Planet. Sci.* **44**, 709–783 (2016).
14. K. D. McKeegan *et al.*, *Science* **332**, 1528–1532 (2011).
15. T. S. Kruijer, C. Burkhardt, G. Budde, T. Kleine, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 6712–6716 (2017).
16. G. Budde *et al.*, *Earth Planet. Sci. Lett.* **454**, 293–303 (2016).
17. P. H. Warren, *Earth Planet. Sci. Lett.* **311**, 93–100 (2011).
18. C. Burkhardt *et al.*, *Earth Planet. Sci. Lett.* **312**, 390–400 (2011).
19. G. M. Poole, M. Rehkämper, B. J. Coles, T. Goldberg, C. L. Smith, *Earth Planet. Sci. Lett.* **473**, 215–226 (2017).
20. Materials and methods are available as supplementary materials.
21. G. Budde, C. Burkhardt, T. Kleine, *Nat. Astron.* **3**, 736–741 (2019).
22. J. A. M. Nanne, F. Nimmo, J. N. Cuzzi, T. Kleine, *Earth Planet. Sci. Lett.* **511**, 44–54 (2019).
23. C. Burkhardt, N. Dauphas, U. Hans, B. Bourdon, T. Kleine, *Geochim. Cosmochim. Acta* **261**, 145–170 (2019).
24. E. Jacquet, F. C. Pignatale, M. Chaussidon, S. Charnoz, *Astrophys. J.* **884**, 32 (2019).
25. Q. R. Shollenberger *et al.*, *Geochim. Cosmochim. Acta* **228**, 62–80 (2018).
26. A. M. Davis, L. Grossman, *Geochim. Cosmochim. Acta* **43**, 1611–1632 (1979).
27. P. J. Sylvester, B. J. Ward, L. Grossman, I. D. Hutcheon, *Geochim. Cosmochim. Acta* **54**, 3491–3508 (1990).
28. G. Budde, T. Kleine, T. S. Kruijer, C. Burkhardt, K. Metzler, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 2886–2891 (2016).
29. A. M. Davis *et al.*, *Geochim. Cosmochim. Acta* **221**, 275–295 (2018).
30. L. Kööp *et al.*, *Nat. Astron.* **2**, 709–713 (2018).
31. A. Trinquier *et al.*, *Science* **324**, 374–376 (2009).
32. J. Zhang, N. Dauphas, A. M. Davis, I. Leya, A. Fedkin, *Nat. Geosci.* **5**, 251–255 (2012).

ACKNOWLEDGMENTS

We thank A. Bischoff for helpful discussions and J. Render and Q. Shollenberger for support and helpful comments on an early draft of this manuscript. We thank N. Marks and B. Jacobsen for assistance with the petrography. We acknowledge the efforts of U. Heitmann for the initial preparation of many of these samples. Allende CAI samples CAI 43 and BB8 were obtained from the collection of the American Museum of Natural History (AMNH) (New York, USA). D. S. Ebel and S. Wallace (Earth and Planetary Sciences, AMNH); R. Matthes, L. Schultz, and L. L. Pryer (Siemens Medical Solutions, Cary, NC); R. Smith (NC Museum of Natural Sciences); and J. Thorstenson (Duke University) are thanked for their support and for providing access to computerized tomography scanners. **Funding:** G.A.B. was supported by a Sofja Kovalevskaja award from the Alexander von Humboldt Foundation. The Deutsche Forschungsgemeinschaft as part of the Research Priority Program SPP 1385 supported T.S.K. (grant KR 4463/1) and T.K. (grant KL 1857/4). An Annette Kade Fellowship (to G.B.) supported characterization of two CAI samples. This study was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344 with release number LLNL-JRNL-802017. **Author contributions:** G.A.B., C.B., and G.B. analyzed samples; G.A.B., C.B., T.S.K., and T.K. devised the project; and all authors contributed to interpretation of the data and writing the manuscript. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** All data are available in tables S1 to S6. Newly measured CAIs were taken from fragments of the meteorites Allende, NWA 6870, and NWA 6717, all of which are curated at the Institut für Planetologie, University of Münster, Germany, except samples CAI 43 and BB8, which were obtained from a fragment of Allende curated at the AMNH, New York, USA. The CAI samples were consumed during the experiments.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/837/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S7
Tables S1 to S6
References (*33*–*63*)

14 October 2019; accepted 16 September 2020
10.1126/science.aaz8482

OPTOMECHANICS

Nano-acoustic resonator with ultralong phonon lifetime

Gregory S. MacCabe[1,2,3]*, Hengjiang Ren[1,2]*, Jie Luo[1,2]†, Justin D. Cohen[1,2]‡, Hengyun Zhou[1,2]§, Alp Sipahigil[1,2], Mohammad Mirhosseini[1,2], Oskar Painter[1,2,3]#

The energy damping time in a mechanical resonator is critical to many precision metrology applications, such as timekeeping and force measurements. We present measurements of the phonon lifetime of a microwave-frequency, nanoscale silicon acoustic cavity incorporating a phononic bandgap acoustic shield. Using pulsed laser light to excite a colocalized optical mode of the cavity, we measured the internal acoustic modes with single-phonon sensitivity down to millikelvin temperatures, yielding a phonon lifetime of up to $\tau_{ph,0} \approx 1.5$ seconds (quality factor $Q = 5 \times 10^{10}$) and a coherence time of $\tau_{coh,0} \approx 130$ microseconds for bandgap-shielded cavities. These acoustically engineered nanoscale structures provide a window into the material origins of quantum noise and have potential applications ranging from tests of various collapse models of quantum mechanics to miniature quantum memory elements in hybrid superconducting quantum circuits.

In optics, geometric structuring at the nanoscale has become a powerful method for modifying the electromagnetic properties of a bulk material, leading to metamaterials capable of manipulating light in unprecedented ways (*1*). In the most extreme case, photonic bandgaps can emerge in which light is forbidden from propagating, markedly altering the emission of light from within such materials (*2*). More recently, a similar phononics revolution (*3*) in the engineering of acoustic waves has led to a variety of new devices, from thermal crystals for controlling the flow of heat (*3*) to phononic topological insulators for scattering-free transport of acoustic waves (*4*).

Phononic bandgap structures, similar to their electromagnetic counterparts, can be used to modify the emission or scattering of phonons. These ideas have recently been explored in quantum optomechanics (*5*–*8*) and electromechanics (*9*) experiments to greatly reduce the mechanical coupling to the thermal environment through acoustic radiation. To date, far less attention has been paid to the impact of geometry and phononic bandgaps on acoustic material absorption and related noise effects (*10*, *11*). Fundamental limits to sound absorption in solids are known to result from the anharmonicity of the host crystal lattice (*12*, *13*). At low temperatures T, in the Landau-Rumer regime ($\omega\tau_{th} \gg 1$), where the thermal phonon relaxation rate (τ_{th}^{-1}) is much smaller than the acoustic frequency (ω), a quantum model of three-phonon scattering can be used to describe phonon-phonon mixing that results in damping and thermalization of acoustic modes (*12*, *13*). Landau-Rumer damping scales approximately as T^{α}, where $\alpha \approx 4$ depends upon the phonon dispersion and density of states (DOS) (*13*). At the very lowest lattice temperatures ($\lesssim$10 K), where Landau-Rumer damping has dropped off, a residual damping emerges due to material defects. These two-level system (TLS) defects (*14*), typically found in amorphous materials, correspond to a pair of nearly degenerate local arrangements of atoms in the solid, which can have both an electric and an acoustic transition dipole, and couple to both electric and strain fields. Recent theoretical analysis shows that TLS interactions with acoustic waves can be substantially altered in a structured material (*10*).

Here we explore the limits of acoustic damping and coherence of a microwave-frequency acoustic nanocavity with a phononic crystal shield that possesses a wide bandgap for all polarizations of acoustic waves. Our nanocavity, formed from an optomechanical crystal (OMC) nanobeam resonator (*5*), supports an acoustic breathing mode at $\omega_m/2\pi \approx 5$ GHz and a colocalized optical resonant mode at $\omega_c/2\pi \approx 195$ THz ($\lambda_c \approx 1550$ nm), which allows us to excite and read out mechanical motion using radiation pressure from a pulsed laser source. This minimally invasive pulsed measurement technique avoids a slew of parasitic damping effects—typically associated with electrode materials and mechanical contact (*15*), or probe fields for continuous readout—and allows for the sensitive measurement of motion at the single-phonon level (*16*). The results of acoustic ringdown measurements at millikelvin temperatures show that damping due to radiation is effectively suppressed by the phononic shield, with breathing-mode quality factors reaching $Q = 4.9 \times 10^{10}$, corresponding to a frequency-Q product of $f - Q = 2.6 \times 10^{20}$.

The devices studied in this work are fabricated from the 220-nm device layer of a silicon-on-insulator (SOI) microchip. [See (*17*) for materials and methods.] In Fig. 1, A and B, we show scanning electron microscope (SEM) images of a single fabricated device, which consists of a coupling optical waveguide, a pair of nanobeam OMC cavities that support both a microwave acoustic and an optical resonant mode, and the acoustic shield that connects the cavities to the surrounding chip substrate. Figure 1C shows finite-element method (FEM) simulations of the microwave acoustic breathing mode and fundamental optical mode of the nanobeam cavity. The design of the OMC cavities, detailed in (*5*), provide strong localization and overlap of the breathing mode and the fundamental optical mode, resulting in a vacuum optomechanical coupling rate (*5*) between photons and phonons of $g_0/2\pi \approx 1$ MHz.

To minimize mechanical clamping losses, the nanobeam is anchored to the Si bulk with a periodic cross structure (*5*). FEM numerical modeling indicates that the cross shield must be carefully designed to provide protection against nanometer-scale disorder, which is inherently introduced during device fabrication (see materials and methods, section II). As shown in Fig. 1, D to F, by including the finite radius of curvature (35 nm) of the corners of the cross shape present in fabricated structures, we could increase the acoustic bandgap to a 3-GHz bandwidth and accurately align the breathing acoustic mode of the nanobeam to the center of the shield gap. Reevaluation of devices measured in (*16*), which did not factor in the finite radius curvature of the fabricated cross shield, indicate a misalignment of the acoustic bandgap with the nanobeam mode.

To investigate the efficacy of the newly designed acoustic shielding in practice, we fabricated and characterized arrays of devices with a scaling of the cross period number from $N_{shield} = 0$ to 10, with all other design parameters held constant. Optical measurements of the acoustic properties of the OMC cavities are performed at millikelvin temperatures in a dilution refrigerator, with a lensed optical fiber used to couple light into and out of each device (*16*). In a first set of measurements of acoustic energy damping, we use a single pulsed laser scheme to perform both excitation and readout of the breathing mode. In this scenario, depicted in Fig. 2, A and B, the laser frequency (ω_l) is tuned to the red motional sideband of the OMC cavity optical resonance, $\Delta \equiv \omega_c - \omega_l \approx +\omega_m$, and is pulsed on for a duration T_{pulse} and then off for a variable time τ. This produces a periodic train of photon pulses

[1]Kavli Nanoscience Institute, California Institute of Technology, Pasadena, CA 91125, USA. [2]Institute for Quantum Information and Matter and Thomas J. Watson, Sr., Laboratory of Applied Physics, California Institute of Technology, Pasadena, CA 91125, USA. [3]AWS Center for Quantum Computing, Pasadena, CA 91125, USA.
*These authors contributed equally to this work.
†Present address: Advanced Quantum Testbed, National Energy Research Scientific Computing Center, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. ‡Present Address: Booz Allen Hamilton Inc, McLean, VA 22102, USA. §Present Address: Department of Physics, Harvard University, Cambridge, MA 02138, USA.
#Corresponding author. Email: opainter@caltech.edu

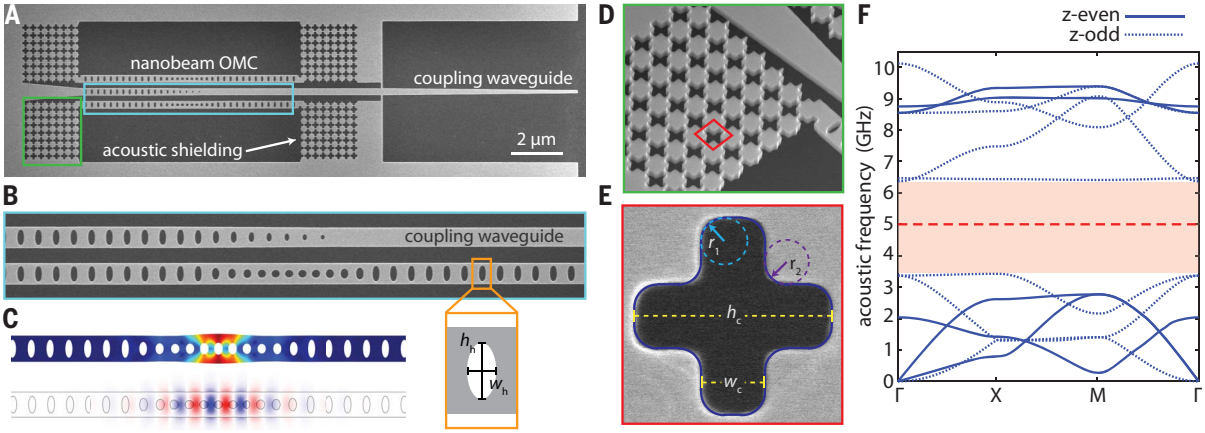


Fig. 1. Nanobeam optomechanical crystal and phononic shield design. (**A**) Scanning electron microscope (SEM) image of a full nanobeam optomechanical crystal (OMC) device fabricated on SOI with $N = 7$ periods of acoustic shielding. A central coupling waveguide allows for fiber-to-chip optical coupling as well as side-coupling to individual nanobeam OMC cavities. (**B**) SEM image of an individual nanobeam OMC and the coupling waveguide, with enlarged illustration of an individual unit cell in the end-mirror portion of the nanobeam. (**C**) FEM simulations of the mechanical (top; total displacement) and optical (bottom; transverse electric field) modes of interest in the nanobeam. Distortion of the mechanical displacement profile is exaggerated for clarity. (**D**) SEM image showing the nanobeam clamping geometry. (**E**) SEM image of an individual unit cell of the cross-crystal acoustic shield. The dashed lines show fitted geometric parameters used in simulation, including cross height ($h_c = 474$ nm), cross width ($w_c = 164$ nm), inner fillet radius (r_1), and outer fillet radius (r_2). (**F**) Simulated acoustic band structure of the realized cross-crystal shield unit cell, with the full acoustic bandgap highlighted in pink. Solid (dotted) lines correspond to modes of even (odd) symmetry in the direction normal to the plane of the unit cell. The dashed red line indicates the mechanical breathing-mode frequency at $\omega_m/2\pi = 5.0$ GHz.

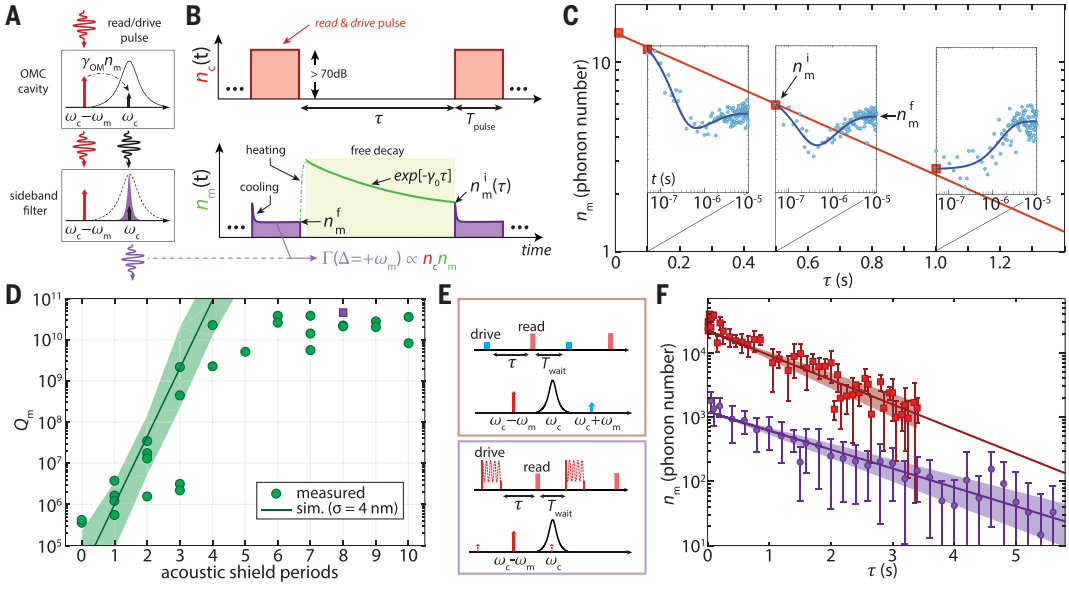


Fig. 2. Ringdown measurements of the acoustic breathing mode. (**A**) Illustration showing the red-detuned ($\Delta = +\omega_m$) optical drive and sideband optical filtering used to realize effective phonon counting measurement. (**B**) Illustration of the ringdown measurement performed using a single pulsed laser for both excitation and readout. Top (bottom): Intracavity photons (phonons) versus time. (**C**) Ringdown measurements of a seven-shield device (device C) for readout pulse amplitude of $n_c = 320$. The series of inset panels show the measured (and fit; solid blue curve) anti-Stokes signal during the optical pulse at a series of pulse delays. (**D**) Plot of the measured breathing mode Q-factor versus number of acoustic shield periods N_{shield}. The solid green line is a fit to the corresponding simulated radiation-limited Q-factor for devices with standard deviation (SD) $\sigma = 4$ nm disorder in hole position and size, similar to the value measured from device SEM image analysis. The shaded green region corresponds to the range of simulated Q values (ensemble size 10) within 1 SD of the mean. The square purple data point represents the measured Q in (F). (**E**) Acoustic excitation is performed coherently by using either a blue-detuned pump (upper diagram) to drive the breathing mode into self-oscillation, or using a radiofrequency-modulated red-detuned pump (*18*) (lower diagram). See materials and methods, section V, for details of the coherent excitation and readout parameters. (**F**) Ringdown measurements performed on an eight-shield device (device D) at large phonon amplitude. For blue-detuned driving (red squares), the fit decay rate is $\gamma_0/2\pi = (0.122 \pm 0.020)$ Hz. For modulated-pump driving (purple circles), the fit decay rate is $\gamma_0/2\pi = (0.108 \pm 0.006)$ Hz. The error bars are 90% confidence intervals of the measured values of n_m^i. The shaded regions are the 90% confidence intervals for the exponential fit curves.

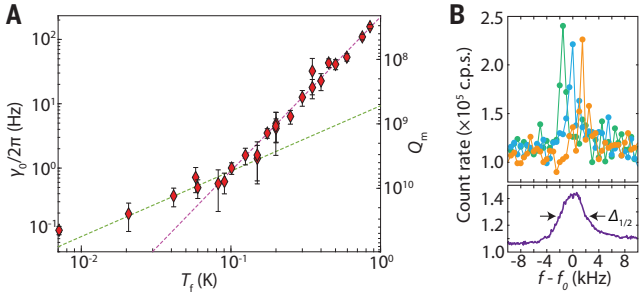


Fig. 3. Temperature dependence of acoustic damping and frequency jitter. (**A**) Plot of the measured breathing-mode energy damping rate, $\gamma_0/2\pi$, as a function of fridge temperature (T_f). Dashed green (magenta) curve is a fit with temperature dependence $\gamma_0 \sim T_f^{1.01}$ ($\gamma_0 \sim T_f^{2.39}$). Error bars are 90% confidence intervals of the exponential fit to measured ringdown curves. (**B**) Two-tone coherent spectroscopy signal. Upper plot: Three individual spectra of rapid frequency sweeps with a frequency step size of 500 Hz and dwell time of 1 ms (resolution bandwidth RBW ≈0.5 kHz). Lower plot: Average spectrum of rapid-scan spectra taken over minutes, showing broadened acoustic response with FWHM linewidth of $\Delta_{1/2}/2\pi = 4.05$ kHz. The large on-resonance response corresponds to an estimated optomechanical cooperativity of $C \equiv \gamma_{OM}/(\gamma_0 + \gamma_p) \gtrsim 1.1$, consistent with the predicted magnitude of back-action damping $\gamma_{OM}/2\pi \approx 817$ Hz and bath-induced damping $\gamma_p/2\pi \approx 120$ Hz at the measurement pump power level $n_c = 0.1$. Data presented in (A) and (B) are for device D.

due to anti-Stokes scattering of the probe laser that are on-resonance with the optical cavity. The anti-Stokes scattered photons are filtered from the probe laser and sent to a single photon detector, producing a photon count rate proportional to the number of phonons in the acoustic resonator (see materials and methods, section III). The resulting intracavity phonon pulse shape, illustrated in Fig. 2B, has a shape determined by a competition between optical absorption–induced heating and optomechanical back-action cooling (*16*). The net result is that the acoustic mode is excited to a small thermal population at the end of the readout pulse, with free decay of the phonon number amplitude in between pulses (see materials and methods, section IV). Plotting the initial mode occupancy at the beginning of the readout pulse (n_m^i) versus delay time τ between pulses results in a ringdown curve of the stored phonon number in the breathing mode, an example of which is displayed in Fig. 2C for a device with $N_{shield} = 7$.

Performing a series of ringdown measurements over a range of devices with varying N_{shield}, and fitting an exponential decay curve to each ringdown, we produce the Q-factor plot in Fig. 2D. We observe an initial trend in Q-factor versus shield number that rises on average exponentially with each additional shield period (×5.5 per period, in good correspondence with the FEM modeling) and then saturates for $N_{shield} \geq 5$ to $Q_m \gtrsim 10^{10}$. As indicated in Fig. 2C, these Q values correspond to ringdown of small, near-single-phonon–level amplitudes. We also perform ringdown measurements at high phonon amplitude using two additional methods displayed schematically in Fig. 2E (see materials and methods, section V). These methods use two laser tones to selectively excite the acoustic breathing mode using a ×1000 weaker excitation and readout optical pulse amplitude ($n_c \lesssim 0.3$). The measured ringdown curves, displayed in Fig. 2F, show the decay from initial phonon occupancies of 10^3 to 10^4 of an eight-shield device (device D; square purple data point in Fig. 2D). The two methods yield similar breathing-mode energy decay rates of $\gamma_0/2\pi = 0.108$ Hz and 0.122 Hz, the smaller of which corresponds to a Q-value of $Q_m = 4.92^{+0.39}_{-0.26} \times 10^{10}$ and a phonon lifetime of $\tau_{ph,0} = 1.47^{+0.09}_{-0.08}$ s. Comparing all three excitation methods with widely varying optical absorption–heating and phonon amplitude, we consistently measure $Q_m \gtrsim 10^{10}$ for devices with $N_{shield} \geq 5$. This represents an improvement by more than three orders of magnitude compared to previous measurements of devices with nonoptimized acoustic shielding (*16*).

To understand the origin of the residual damping for large N_{shield} we also measured the temperature dependence of the energy damping rate for a number of the high-Q, large N_{shield} devices. In Fig. 3A, we plot the energy damping rate versus fridge temperature (T_f) for the highest-Q eight-shield device (device D), typical of all of these types of devices that we measured, which shows an approximately linear rise in temperature up to $T_f \approx 100$ mK and then a much faster $\sim (T_f)^{2.4}$ rise in the damping above this temperature. This rather anomalous temperature scaling of the damping does not fit the standard damping models relating to three-phonon scattering, and estimates of the magnitude of Landau-Rumer damping of the breathing mode indicate that three-phonon scattering in Si is far too weak at $T_f \lesssim 1$ K to explain the absolute level of measured damping (see materials and methods, section VI). Canonical models of TLS-induced damping also fail to predict the measured temperature scaling of the damping. Analysis of the interactions of TLS in the surface of the silicon with the localized acoustic modes in the confined geometry of the OMC cavity sructure, however, show that TLS interactions can explain all of the observed breathing-mode behavior when taking into account the highly modified density of phonon states in the shielded OMC cavity (see supplementary text, sections VII and VIII). In particular, we find that damping due to nearly resonant TLS is suppressed owing to the bandgap of the phononic shield and that relaxation damping from nonresonant TLS can explain the magnitude and low-temperature dependence of the breathing-mode damping and the lack of saturation of the damping with both temperature and acoustic amplitude.

In addition to energy damping, the coherence of the high-Q breathing modes is also of interest. Using the same pump-probe excitation method (*18*) as in the ringdown measurements of Fig. 2, E and F, we can measure the breathing-mode acoustic spectrum by sweeping the probe tone across the acoustic resonance. Figure 3B shows the measured acoustic spectrum of device D at the fridge base temperature $T_f = 7$ mK and a low pump power of $n_c = 0.1$. The top plot shows a series of rapid spectral scans (40 ms per scan), which captures a jittering narrow acoustic line. Scanning more slowly than the 40-ms per scan begins to introduce broadening into the single-scan spectral line, providing a rough estimate for the (inverse) bandwidth of the frequency jitter noise of the breathing mode. An ensemble average of these scans, taken over several minutes, yields a broadened acoustic full width at half-maximum (FWHM) linewidth of $\Delta_{1/2}/2\pi = 4.05$ kHz (bottom plot of Fig. 3B). This behavior is characteristic of coherent-like frequency noise, such as flicker or low-frequency telegraph noise. In this case, the coherence time is given by $\tau_{coh,0} \approx 4(\ln(2))^{1/2}/\Delta_{1/2} \approx 130$ μs (*19*, *20*). The likely source of this noise is again fluctuating TLS (see supplementary text, section VIII).

Several subtle points regarding the energy decay and coherence measurements should be noted. First, the breathing-mode spectral linewidth (and thus coherence time) is measured in the presence of a weak laser pump field, in contrast to energy decay measurements, in which the laser is pulsed off. Although we did not observe a power dependence of the measured time-average acoustic linewidth, the influence of the laser field on the acoustic mode coherence time cannot be ruled out. Second, the pump-probe technique used to measure the acoustic spectrum excites the breathing mode to an estimated mode occupancy of order 1000; of relevance to many quantum applications is the single-phonon coherence time.

With the recent demonstration of strong dispersive coupling of a superconducting qubit to similar nanomechanical acoustic cavities (*21*), measurement of single-phonon coherence time in the absence of light fields should be possible. Additionally, the low-frequency character of the frequency noise measured in this work indicates that the acoustic coherence time may be substantially increased toward the energy decay time through techniques such as dynamic decoupling (*22*). Finally, the attribution of the residual damping and the frequency jitter noise to surface TLS indicates that further study of the silicon surface composition may lead to a better understanding of the microscopic nature of the TLS and help identify further surface preparation methods for their removal (see materials and methods, section I, and supplementary text, section VIII).

More generally, the results presented here show that the advanced methods of nanofabrication and cavity optomechanics provide a new toolkit to explore quantum acoustodynamics in solid-state materials (*23*). For example, continued studies of phononic engineered structures in the quantum regime may lead to techniques for modifying the behavior of quasiparticles in superconductors (*24*), new approaches for mitigating decoherence (noise) in superconducting (*25*) and color center (*26*) qubits, and even new TLS–acoustic cavity qubit architectures (*27*). The extremely small motional mass [$m_{\mathrm{eff}} = 136$ fg (*5*)] and narrow linewidth of the OMC cavity also make it ideal for precision mass sensing (*28*) and in exploring limits to alternative quantum collapse models (*29*). Perhaps most intriguing is the possibility of creating a hybrid quantum architecture consisting of acoustic and superconducting quantum circuits (*21*, *30*), where the small scale, reduced cross-talk, and ultralong coherence time of quantum acoustic devices may provide substantial improvements in connectivity and performance of current quantum hardware.

REFERENCES AND NOTES

1. J. D. Joannopoulos, S. G. Johnson, J. N. Winn, R. D. Meade, *Photonic Crystals: Molding the Flow of Light* (Princeton Univ. Press, ed. 2nd, 2008).
2. S. John, J. Wang, *Phys. Rev. B Condens. Matter* **43**, 12772–12789 (1991).
3. M. Maldovan, *Nature* **503**, 209–217 (2013).
4. J. Cha, K. W. Kim, C. Daraio, *Nature* **564**, 229–233 (2018).
5. J. Chan, A. H. Safavi-Naeini, J. T. Hill, S. Meenehan, O. Painter, *Appl. Phys. Lett.* **101**, 081115 (2012).
6. P.-L. Yu *et al.*, *Appl. Phys. Lett.* **104**, 023510 (2014).
7. Y. Tsaturyan, A. Barg, E. S. Polzik, A. Schliesser, *Nat. Nanotechnol.* **12**, 776–783 (2017).
8. A. H. Ghadimi *et al.*, *Science* **360**, 764–768 (2018).
9. M. Kalaee *et al.*, *Nat. Nanotechnol.* **14**, 334–339 (2019).
10. R. O. Behunin, F. Intravaia, P. T. Rakich, *Phys. Rev. B* **93**, 224110 (2016).
11. B. D. Hauer, P. H. Kim, C. Doolin, F. Souris, J. P. Davis, *Phys. Rev. B* **98**, 214303 (2018).
12. L. Landau, G. Rumer, *Phys. Z. Sowjet.* **11**, 18 (1937).
13. G. P. Srivastava, *The Physics of Phonons* (Taylor & Francis, 1990).
14. W. A. Phillips, *Rep. Prog. Phys.* **50**, 1657–1708 (1987).
15. S. Galliou *et al.*, *Sci. Rep.* **3**, 2132 (2013).
16. S. M. Meenehan *et al.*, *Phys. Rev. X* **5**, 041002 (2015).
17. See supplementary materials.
18. A. H. Safavi-Naeini *et al.*, *Nature* **472**, 69–73 (2011).
19. G. Di Domenico, S. Schilt, P. Thomann, *Appl. Opt.* **49**, 4801–4807 (2010).
20. P. J. J. O'Malley, thesis, University of California, Santa Barbara (2016).
21. P. Arrangoiz-Arriola *et al.*, *Phys. Rev. X* **8**, 031007 (2018).
22. J. Bylander *et al.*, *Nat. Phys.* **7**, 565–570 (2011).
23. W. H. Renninger, P. Kharel, R. O. Behunin, P. T. Rakich, *Nat. Phys.* **14**, 601–607 (2018).
24. K. Rostem, P. J. de Visser, E. J. Wollack, *Phys. Rev. B* **98**, 014522 (2018).
25. J. M. Martinis, A. Megrant, UCSB final report for the CSQ program: Review of decoherence and materials physics for superconducting qubits. *arXiv:1410.5793* (2014).
26. Y.-I. Sohn *et al.*, *Nat. Commun.* **9**, 2012 (2018).
27. T. Ramos, V. Sudhir, K. Stannigel, P. Zoller, T. J. Kippenberg, *Phys. Rev. Lett.* **110**, 193602 (2013).
28. M. S. Hanay *et al.*, *Nat. Nanotechnol.* **10**, 339–344 (2015).
29. S. Nimmrichter, K. Hornberger, K. Hammerer, *Phys. Rev. Lett.* **113**, 020405 (2014).
30. M. H. Devoret, R. J. Schoelkopf, *Science* **339**, 1169–1174 (2013).
31. G. S. MacCabe, H. Ren, J. Luo, J. D. Cohen, H. Zhou, A. Sipahigil, M. Mirhosseini, O. Painter, Data for "Nano-acoustic resonator with ultralong phonon lifetime," Version 1, Zenodo (2020); http://doi.org/.

ACKNOWLEDGMENTS

Funding: This work was supported by the ARO Quantum Opto-Mechanics with Atoms and Nanostructured Diamond MURI program (grant N00014-15-1-2761), the ARO-LPS Cross-Quantum Systems Science and Technology program (grant W911NF-18-1-0103), the Institute for Quantum Information and Matter, an NSF Physics Frontiers Center (grant PHY-1733907) with support of the Gordon and Betty Moore Foundation, and the Kavli Nanoscience Institute at Caltech. H.R. gratefully acknowledges support from the National Science Scholarship from A*STAR, Singapore. **Author contributions:** G.S.M., H.R., J.D.C., and O.P. came up with the concept and planned the experiment. G.S.M., H.R., J.L., H.Z., and J.D.C. performed the device design and fabrication. G.S.M. and H.R. performed the measurements. G.S.M., H.R., J.L., A.S., M.M., and O.P. analyzed the data. All authors contributed to the writing of the manuscript. **Competing interests:** G.S.M, H.R., J.L., A.S., M.M., and O.P. acknowledge two related patent applications that draw on the work presented in this article: U.S. Patent Application no. 16/293,457, "Techniques for Bidirectional Transduction of Quantum Level Signals Between Optical and Microwave Frequencies Using a Common Acoustic Intermediary"; U.S. Patent Application no. 16/293,455, "Techniques for Transduction and Storage of Quantum Level Signals". J.L. is also affiliated with Anyon Computing Inc., Pasadena, CA 91101, USA. **Data and materials availability:** Data and data analysis code are available through Zenodo (*31*). All other data that support the findings of this study are in the main text and supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/840/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S19
Tables S1 to S4
References (*32*–*56*)

10 May 2020; accepted 5 October 2020
10.1126/science.abc7312

NEUROSCIENCE

A thalamocortical top-down circuit for associative memory

M. Belén Pardi[1], Johanna Vogenstahl[1], Tamas Dalmay[1,2], Teresa Spanò[1,3], De-Lin Pu[1], Laura B. Naumann[4,5], Friedrich Kretschmer[1], Henning Sprekeler[4,5], Johannes J. Letzkus[1,6]*

The sensory neocortex is a critical substrate for memory. Despite its strong connection with the thalamus, the role of direct thalamocortical communication in memory remains elusive. We performed chronic in vivo two-photon calcium imaging of thalamic synapses in mouse auditory cortex layer 1, a major locus of cortical associations. Combined with optogenetics, viral tracing, whole-cell recording, and computational modeling, we find that the higher-order thalamus is required for associative learning and transmits memory-related information that closely correlates with acquired behavioral relevance. In turn, these signals are tightly and dynamically controlled by local presynaptic inhibition. Our results not only identify the higher-order thalamus as a highly plastic source of cortical top-down information but also reveal a level of computational flexibility in layer 1 that goes far beyond hard-wired connectivity.

Accurate perception of the world requires integration of external sensory signals with internally generated information representing the previously acquired relevance of stimuli (*1*, *2*). Whereas decades of work have elucidated how the sensory neocortex processes physical stimulus features, encoding of memory-related top-down information by brain-wide afferents remains poorly understood. Neocortical layer 1 (L1) receives a range of top-down inputs from diverse sources (*1*, *3*) and could therefore be a key node for integration of these signals. One prominent yet little understood projection to L1 arises from the higher-order thalamus (*4*–*6*). We hypothesized that these afferents constitute a source of experience-dependent top-down information, distinct from the classical bottom-up encoding that has been widely observed in first-order thalamic nuclei (*7*).

To investigate the function of thalamocortical communication in memory, we focused on afferents to the auditory cortex, a critical area for associative memory (*8*). The higher-order auditory thalamus [HO–medial geniculate (HO-MG)] comprises several nuclei that surround the first-order auditory thalamus [ventral division of the MG (MGv)] like a shell. To selectively target protein expression to the HO-MG while avoiding the MGv, we used conditional adeno-associated viral vector (AAV)–mediated expression in CR-IRES-Cre mice (*9*) (Fig. 1, A and B, and fig. S1; for full results of all statistical tests in this study, see supplementary text section of the supplementary materials). Anterograde tracing revealed that HO-MG axons display a mediotemporal density gradient that closely matches the location of cortical areas critical for associative memory (*8*), with the highest density in temporal cortical areas, including the secondary auditory cortex (AuV) and the temporal association cortex (Fig. 1C). We therefore focused on the most temporal area of the auditory cortex (i.e., AuV), where axons were strongly enriched in upper L1 (Fig. 1C).

To identify the postsynaptic targets of HO-MG afferents in L1, we used channelrhodopsin-2 (ChR2)–assisted circuit mapping in acute brain slices (*10*) while recording from L1 interneurons (L1 INs) and pyramidal neurons (Pyrs) in L2/3 and L5 (Fig. 1, D to F). Local light stimulation in L1 under action potential block (fig. S2, A and B) excited the three neuron types with similar amplitudes (Fig. 1, E and F), albeit with different probability (fig. S2C). In line with synaptic contacts residing on distal dendrites of Pyrs, excitatory currents recorded in L5 displayed longer latencies (Fig. 1F). These results indicate that HO-MG L1 input provides

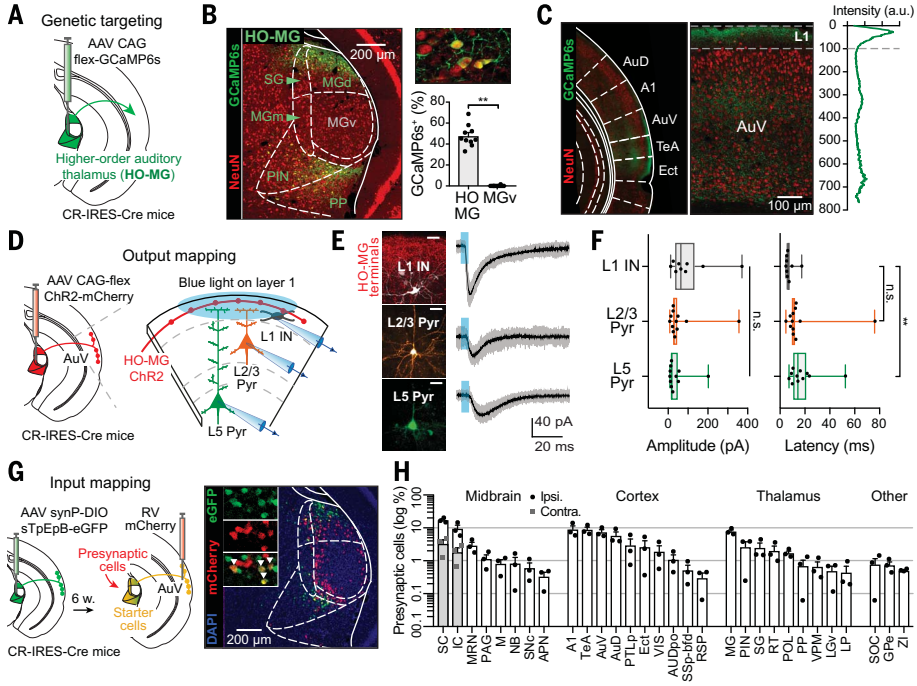


Fig. 1. The higher-order auditory thalamus provides major top-down input to cortical layer 1. (**A**) Experimental strategy for targeting of HO-MG in calretinin (CR)–IRES-Cre mice. (**B** and **C**) Immunostaining against GCaMP6s and neuronal marker NeuN. (B) Auditory thalamus, posterior intralaminar nucleus (PIN) detail, and quantification of $NeuN^+$ cells expressing GCaMP6s (Wilcoxon test, n = 10 slices, 5 mice). (C) HO-MG axons in cortex and mean fluorescence intensity (arbitrary units, a.u.) versus distance from pia (in micrometers, n = 3 slices, 3 mice). (**D**) ChR2-assisted mapping of connections between HO-MG axons in AuV and L1 INs and L2/3 and L5 Pyrs in acute slices. Light stimulation (5 ms) was restricted to L1. (**E**) Recorded cells and $ChR2^+$ HO-MG axons in L1 (pseudocolored; scale bars, 20 µm), and light-evoked excitatory postsynaptic currents (gray, single trials; black, mean; blue bar, light). (**F**) Current amplitude and latency to peak (quartiles, median and range, Kruskal-Wallis test with Dunn's multiple comparisons test for latency, n = 9 L1 INs, n = 12 L2/3 Pyrs, n = 11 L5 Pyrs). (**G**) Transsynaptic rabies (RV) tracing of inputs to HO-MG neurons that project to AuV. Glycoprotein, EnvA receptor (TVA), and enhanced green fluorescent protein (eGFP) were conditionally expressed in the HO-MG. RV-mCherry was injected into the auditory cortex 6 weeks later, leading to retrograde transduction of starter cells. Right: Representative expression in the auditory thalamus [4′,6-diamidino-2-phenylindole (DAPI) counterstain]. Inset from PIN: eGFP, mCherry, and overlay; arrowheads indicate starter cells. (**H**) Distribution of presynaptic cells ipsilateral and contralateral to injection, normalized per animal (logarithmic scale, n = 3 mice). Abbreviations used in (B), (C), and (H) are defined in table S1. Data shown as mean ± SEM, except in (F). n.s. $P > 0.05$, $^{**}P < 0.01$.

widespread control over cortical activity that can affect circuit processing in diverse ways: Recruitment of L1 INs can generate disinhibition of lower layers that is particularly relevant for associative learning (*11*), and it can also directly inhibit distal dendrites (*3*). Furthermore, excitation of Pyr distal dendrites can trigger dendritic spikes and facilitate long-term potentiation (*4*, *12*).

These observations are consistent with the hypothesized role of the HO-MG as a major top-down projection. Top-down inputs are predicted to convey feedback information from diverse sources. We therefore identified the brain-wide inputs specifically to auditory cortex–projecting HO-MG neurons, using dual-site-injection Cre-dependent transsynaptic retrograde rabies tracing (*13*, *14*) (Fig. 1G). This experiment revealed diverse input sources (Fig. 1H and fig. S3), including cortical areas, intermediate and deep layers of the superior colliculus, external capsule of the inferior colliculus, and the periaqueductal gray. The HO-MG thus transmits integrated information to cortical L1, in line with previous work on input to the entire HO-MG (*7*, *15*).

To directly test whether HO-MG afferents convey memory-related information to the cortex, we used discriminative threat conditioning (DTC) to endow auditory stimuli with experience-dependent relevance in a controlled manner (*16*). Conditioning was performed by repeatedly pairing one of two initially neutral conditioned stimuli (CSs) with mild foot-shocks. We used trains of frequency-modulated sweeps as CSs [trains are termed CSs, individual sweeps are termed frequency-modulated (FM) sweeps], naturalistic stimuli that critically require cortical processing (*8*, *17*). We combined this approach with two-photon calcium imaging of HO-MG boutons expressing GCaMP6s in L1 of awake, head-fixed mice after identification of AuV by intrinsic imaging (Fig. 2A and fig. S4) (*3*, *8*). We imaged boutons in response to the reinforced CS^+ and the unreinforced CS^- 1 day before (habituation) and 1 day after conditioning (recall 1, Fig. 2A). CSs elicited different levels of threat memory, measured as the change in eye pupil dilation during head fixation (from habituation to recall), and as freezing in freely behaving conditions (recall 2, Fig. 2, A and B, and fig. S5, A to C), readouts that displayed a robust positive correlation (fig. S5D) (*3*, *8*). Only boutons identified in habituation and recall were analyzed (Fig. 2C and fig. S4), enabling an unbiased, longitudinal characterization.

CS^+ stimulation during habituation in awake head-fixed mice elicited clear responses in a subset of boutons (Fig. 2, D and E). However, HO-MG synaptic responses decayed rapidly to baseline after a few trials (Fig. 2F), indicating that stimulus encoding by HO-MG afferents is highly sensitive to novelty. Another distinctive feature was their markedly delayed kinetics, with most of the response occurring after each FM sweep (Fig. 2, D and G)—very different from the short-latency responses typically found in first-order MGv (*7*).

In contrast, CS^+ responses in memory recall displayed diverse learning-related plastic changes leading to a much stronger activation of HO-MG boutons. First, while CS^+ responses of individual boutons either increased or decreased, the total number of responsive boutons grew by

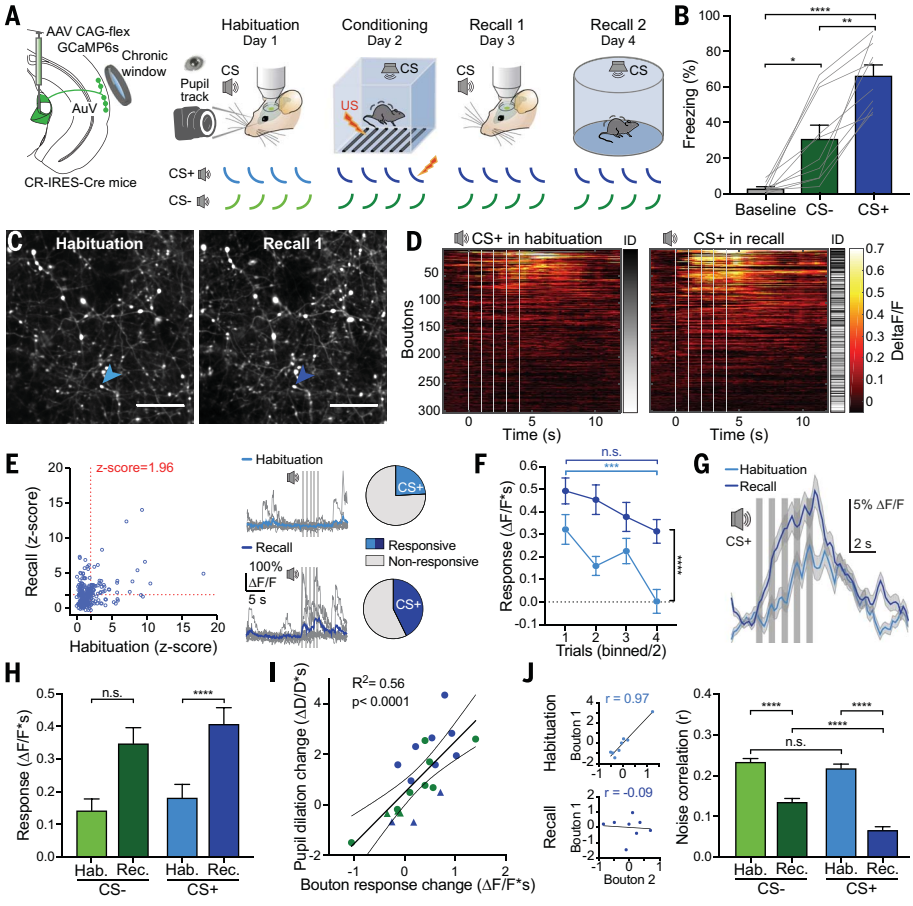


Fig. 2. Experience-dependent signaling by HO-MG afferents to cortical L1. (**A**) Experimental strategy. US, unconditioned stimulus. (**B**) Freezing behavior of learner mice in recall 2 [imaging of these animals is shown in (D) to (H) and (J); one-way repeated measures analysis of variance (RM ANOVA) with Tukey's multiple comparisons test, n = 10 mice]. (**C**) Representative boutons in habituation and recall (scale bars, 20 μm). (**D**) Mean bouton responses sorted by amplitude, and gray scale rank order bouton ID [vertical lines indicate FM-sweep onset; in (D) to (H), n = 302 boutons, 10 mice]. (**E**) Peak z-score of individual bouton CS^+ responses (z-scores > 1.96 are considered responsive). Right: Representative bouton indicated in (C) (gray, single trials; blue, mean) and fraction of responsive boutons (habituation, 73/302; recall, 129/302). ΔF/F, change in fluorescence intensity. (**F**) Response integral during CS^+ over trials (every two binned) in habituation versus recall (two-way RM ANOVA, $P < 0.0001$ for session, with Sidak's multiple comparisons test between trials). (**G**) Mean ± SEM bouton response. (**H**) Response integral during CSs (Friedman test with Dunn's multiple comparisons test). (**I**) Bouton versus pupil response change between habituation and recall in conditioned mice [Pearson correlation, linear fit and 95% confidence bands, n = 22 mean responses from CS^+ blue and CS^- green, eight learner mice (dots) and three nonlearner mice (triangles)]. (**J**) Left: Trial-by-trial noise correlation between response integrals in a pair of boutons (ΔF/F*s, Pearson correlation, linear fit, n = 8 single presentations). Right: Noise correlation between CS responsive and nonresponsive bouton pairs in learner mice (Kruskal-Wallis test with Dunn's multiple comparisons test, CS^+ habituation n = 2212, CS^+ recall n = 3130, CS^- habituation n = 2989, CS^- recall n = 3230 correlations). Data shown as mean ± SEM. n.s. $P > 0.05$, *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$.

[1]Max Planck Institute for Brain Research, 60438 Frankfurt, Germany. [2]Donders Centre for Neuroscience, Faculty of Science, Radboud University, 6525 AJ Nijmegen, Netherlands. [3]Faculty of Biological Sciences, Goethe Universität Frankfurt, 60438 Frankfurt, Germany. [4]Bernstein Center for Computational Neuroscience Berlin, 10115 Berlin, Germany. [5]Department of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany. [6]Institute for Physiology I, Faculty of Medicine, University of Freiburg, 79104 Freiburg, Germany.
***Corresponding author. Email: johannes.letzkus@brain.mpg.de**

77% after learning (Fig. 2, D and E). Second, CS^+ representation was much more persistent over trials than in habituation (Fig. 2F). Third, CS^+ responses became more time-locked to the onset of the CS (Fig. 2, D and G, and fig. S6, A to C) and preceded behavioral memory expression (fig. S6, D to F). Fourth, the overall CS^+ response was strongly potentiated with learning (Fig. 2, G and H, and fig. S7), mainly during the CS but also, to a lesser extent, after CS offset (fig. S7D).

Grouping animals according to the degree of behavioral CS discrimination revealed that boutons from mice with low CS^- freezing displayed no response potentiation to the CS^-. Conversely, potentiation of CS^- responses occurred in animals with strong CS^- threat memory (fig. S8, A and B). Across all mice, the observed changes were highly significant for the CS^+ but not for the CS^- (Fig. 2H, figs. S6B and S7, A to D) and were absent in animals that did not learn the CS association (fig. S9). To directly address the relationship between plasticity and memory strength, we measured the change in pupil dilation from habituation to recall (fig. S5, B to D). Analysis of bouton response changes per mouse revealed that interindividual variability of potentiation magnitude strongly correlated with—and could thus be predictive of—memory strength (Fig. 2I). Analyses from each session separately indicated a significant correlation between absolute bouton and pupil responses only in recall (fig. S5E), further underpinning the impact of associative memory on HO-MG responses.

Finally, we analyzed activity correlations between boutons, which can limit the content of information transfer. Trial-by-trial noise correlations between all pairs of boutons displayed a decrease in memory recall relative to habituation (fig. S7F). This was mediated specifically by a decrease in the correlation between CS responsive and nonresponsive boutons (Fig. 2J and fig. S8D), while correlations among responsive and among nonresponsive boutons were maintained (fig. S7F).

Collectively, these data demonstrate that the HO-MG conveys information about the behavioral relevance of sounds to auditory cortex L1, which depends on previous experience both in the short term, during habituation, and in the long term, after associative conditioning. Whereas remodeling of thalamocortical systems has been described in development and after sensory deprivation (*18*), our data highlight the central role of plasticity in higher-order thalamic afferents for associative memory in the adult.

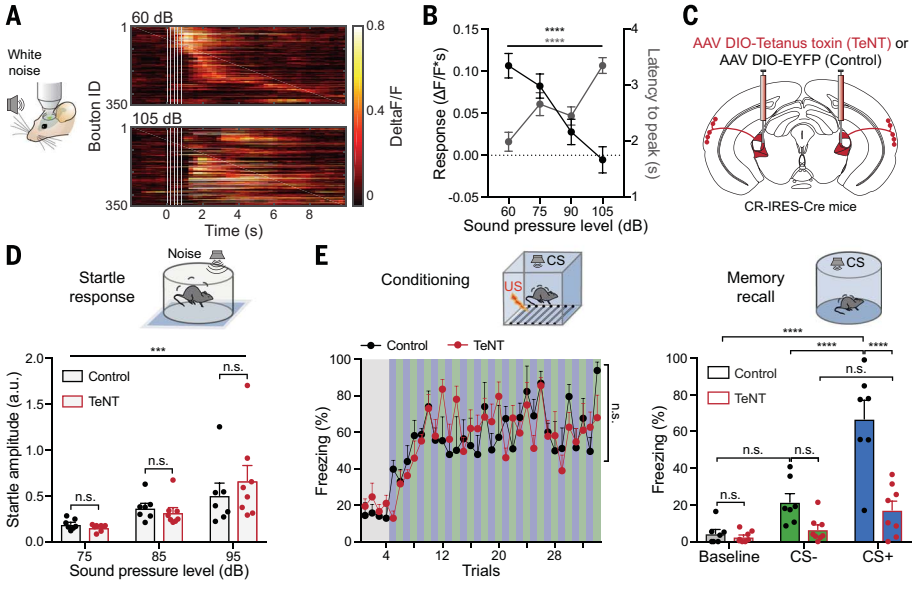


Fig. 3. The HO-MG is required for auditory associative memory. (**A**) HO-MG bouton responses to white noise–trains of 60 and 105 dB sound pressure level (SPL), sorted by amplitude to 60 dB. (**B**) Response integral during noise-trains of increasing intensities (Friedman test, n = 350 boutons, 4 mice) and response latency to peak for responsive boutons (Kruskal-Wallis test, 60 dB n = 139, 75 dB n = 148, 90 dB n = 149, 105 dB n = 225 boutons). (**C**) AAV-mediated bilateral expression of TeNT and enhanced yellow fluorescent protein (EYFP) (control). (**D**) Startle response (two-way RM ANOVA, $P < 0.001$ for sound intensity, Sidak's multiple comparisons test between animal groups, n.s.). In (D) and (E), n = 8 TeNT and n = 7 control mice. (**E**) TeNT inhibition during DTC. Left: Freezing behavior during conditioning (two-way RM ANOVA, n.s. for animal groups; gray, baseline time; blue, CS^+ trials; green, CS^- trials). Right: Freezing behavior during memory recall (two-way RM ANOVA with Tukey's multiple comparisons test between trial types and Sidak's multiple comparisons test between animal groups). Data shown as mean ± SEM. n.s. $P > 0.05$, ***$P < 0.001$, ****$P < 0.0001$.

A second form of salience that is independent of experience derives from the physical properties of sensory stimuli, such as sound intensity (*19*). To address whether HO-MG afferents also encode this bottom-up salience, we imaged boutons in response to noise-train stimuli at different intensities (Fig. 3, A and B). In contrast to both the changes observed after learning and the typical encoding in first-order thalamic afferents (*20*), bouton responses during the noise-train decreased with increasing stimulus intensity (Fig. 3B). This effect was generated by a successive increase in the latency of responses, whereas activity still occurred after noise-train offset (fig. S10), indicating a disengagement of this pathway during stimuli with increasing bottom-up salience.

We next tested the requirement of HO-MG signaling for associative memory. Synaptic silencing of HO-MG neurons with tetanus toxin (TeNT) (*3*) left basic auditory processing addressed by startle responses intact (Fig. 3, C and D) but yielded an impairment of memory expression to the CS^+ along with deficits in discrimination after DTC (Fig. 3E, right, and fig. S11). Conversely, behavior during acquisition and contextual threat recall was not affected (Fig. 3E, left, and fig. S11, D and F), indicating that the HO-MG is critically required for associative auditory memory but dispensable for foot-shock perception and nonauditory memory.

Whereas top-down information is typically considered to be independent of activity in the local circuit (*1*, *2*), we hypothesized that these afferents could be under the functional control of a specialized type of neurogliaform IN that densely innervates L1 and expresses the selective marker NDNF (neuron-derived neurotrophic factor) (*3*). Because these neurogliaform cells can perform γ-aminobutyric acid–mediated (GABAergic) volume transmission (*21*, *22*), they could control the release probability of HO-MG boutons through presynaptic GABA type B receptor ($GABA_BR$)–mediated inhibition (*23*) (Fig. 4A). Such a mechanism would add a layer of complexity to neuronal processing by enabling dynamic and target area–specific control of top-down afferent information.

We first asked whether HO-MG boutons in the cortex can be inhibited through $GABA_BRs$. Recordings in acute brain slices revealed that the $GABA_BR$ agonist baclofen reduced optogenetically evoked postsynaptic current amplitude and abolished short-term depression (Fig. 4, B and C, and fig. S12, A and B), consistent with modulation of release probability by presynaptic $GABA_BRs$. Both effects were reversed by application of the $GABA_BR$ antagonist CGP55845. When axons were stimulated with a naturalistic train of light pulses comprising a range of instantaneous frequencies (Fig. 4D and fig. S12C), baclofen enhanced transmission of higher frequencies, indicating that

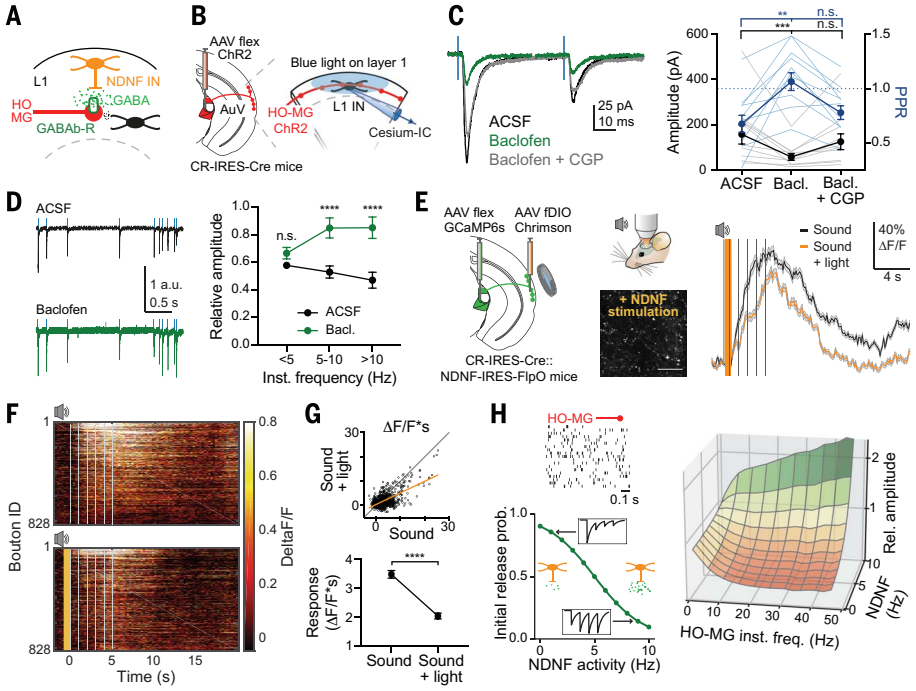


Fig. 4. Local modulation by layer 1 NDNF interneurons through presynaptic $GABA_B$ receptors. (**A**) Hypothesis: NDNF INs in L1 modulate HO-MG afferents through presynaptic $GABA_B$R-mediated inhibition. (**B**) ChR2-expressing HO-MG axon stimulation in AuV L1 in acute slices and recording of L1 INs. (**C**) Left: Representative excitatory postsynaptic currents (EPSCs, 0.5-ms pulses, 20 Hz) in control [artificial cerebrospinal fluid (ACSF)], baclofen, and baclofen+CGP. Right: First pulse response amplitude (Friedman test with Dunn's multiple comparisons test) and paired-pulse ratio (PPR, one-way RM ANOVA with Sidak's multiple comparisons test, n = 10 neurons, 7 mice). (**D**) Left: Representative EPSCs during naturalistic stimulation, normalized to first response. Right: Normalized response amplitude, grouped according to instantaneous frequency (two-way RM ANOVA with Sidak's multiple comparisons between ACSF and baclofen, n = 10 neurons, 4 mice). (**E**) HO-MG boutons imaged in AuV L1 (scale bar, 40 μm) in response to a white noise–train, with optogenetic NDNF L1 IN stimulation in interleaved trials. Traces: Mean ± SEM of boutons [orange bar, light; (E) to (G): n = 828 sound-responsive boutons, 4 mice]. (**F**) Mean individual bouton responses. (**G**) Full response integral. Top: Individual boutons [gray, identity line; orange, linear fit; coefficient of determination (R^2) = 0.40, $P < 0.001$]. Bottom: Mean responses (Wilcoxon test). (**H**) Computational model of NDNF L1 IN modulation of HO-MG afferents. Top left: Representative Poisson spike trains of HO-MG afferents. Bottom left: Dependence of HO-MG bouton initial release probability on NDNF L1 IN activity. Insets: Example EPSCs for constant 25-Hz spike input. Right: Mean EPSC amplitude normalized to the first response as a function of HO-MG instantaneous spike frequency and NDNF L1 IN activity. Data shown as mean ± SEM. n.s. $P > 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

presynaptic $GABA_B$R activation can robustly modulate the frequency content of afferent information.

While these results demonstrate plausibility, a major open question is whether GABA released from endogenous sources in vivo can exert similar control. We capitalized on the genetic tractability of NDNF L1 INs for selective in vivo optogenetic stimulation (*3*) and combined this with imaging of HO-MG boutons (Fig. 4E). NDNF L1 IN excitation caused a robust inhibition of HO-MG synaptic activity (Fig. 4, E to G), whereas light stimulation in control animals had no effect (fig. S13A). Consistent with $GABA_B$R-mediated presynaptic modulation, NDNF L1 IN stimulation also affected temporal response dynamics (fig. S13, B to D). Together, these results reveal that NDNF L1 INs exert tight control over HO-MG afferent information. This interpretation is quantitatively borne out by a model of our data, which highlights the power of this mechanism for dynamic selection of afferent frequency content through local NDNF L1 INs (Fig. 4H and fig. S14). Intriguingly, activity of L1 INs depends on both learning (*3*) and sensory context (*24*), suggesting that presynaptic inhibition may itself be subject to modulation by behavioral state.

Although the higher-order thalamus comprises half of all thalamic nuclei, has selectively expanded in the evolution of higher primates, and has been implicated in attention and other cognitive capacities from rodents to humans (*25*, *26*), the function of its connection to the cortex remains poorly understood. Our results identify the HO-MG as a highly experience-dependent source of top-down information that is critical for associative memory, integrates inputs from a range of brain areas, connects widely to cortical circuit elements, and whose information transfer is in turn modulated by local inhibition in L1. Notably, bottom-up and top-down salience recruit distinct coding mechanisms in HO-MG afferents to L1: Whereas CSs elicited potentiated and more time-locked responses after learning, increased stimulus intensity had the opposite effect, reducing and delaying the signals. These results suggest that signaling within L1 may be particularly relevant for perception of stimuli with low bottom-up salience, in line with the central role of activity in the apical dendrites of L5 Pyrs for perception at the detection threshold (*27*). Given the described functional differences of HO-MG nuclei (*7*, *28*), future work is needed to address the potentially distinct contribution of their projections to the auditory cortex.

In L1, dendrites of Pyrs are a central substrate for integration and compartmentalization of diverse afferents (*12*). While our results indicate that these mechanisms will contribute to processing of HO-MG inputs (*4*), we also identify a previously unknown form of modulation of afferent information acting on presynaptic $GABA_B$Rs, which can dynamically control synaptic transmission. The dependence on local inhibition not only implies that information encoded by long-range boutons can differ from that encoded at the level of the soma but also that collaterals from an individual axon may convey distinct information to different target brain areas. Given that presynaptic $GABA_B$Rs are present on a wide variety of synapses (*23*, *29*), it is likely that additional afferents to cortical L1 are subject to this form of modulation.

Beyond memory expression, top-down afferents are involved in a range of additional functions, such as credit assignment during learning (*30*) and predictive coding, in which a generative model of the environment is fed back to lower processing stations in order to encode mismatches between expected and actual sensory inputs (*1*, *2*). The present multidisciplinary dissection of cortical HO-MG afferents in memory-related top-down encoding will therefore also enable future work toward a better mechanistic understanding of these diverse cognitive processes.

REFERENCES AND NOTES

1. A. M. Bastos *et al.*, *Neuron* **76**, 695–711 (2012).
2. G. B. Keller, T. D. Mrsic-Flogel, *Neuron* **100**, 424–435 (2018).
3. E. Abs *et al.*, *Neuron* **100**, 684–699.e6 (2018).
4. F. Gambino *et al.*, *Nature* **515**, 116–119 (2014).
5. M. M. Roth *et al.*, *Nat. Neurosci.* **19**, 299–307 (2016).
6. N. J. Audette, S. M. Bernhard, A. Ray, L. T. Stewart, A. L. Barth, *Neuron* **103**, 277–291.e4 (2019).
7. J. A. Winer, in *The Auditory Cortex*, J. A. Winer, C. E. Schreiner, Eds. (Springer, 2011), pp. 41–74.
8. T. Dalmay *et al.*, *Neuron* **104**, 1180–1194.e7 (2019).

9. E. Lu, D. A. Llano, S. M. Sherman, *Hear. Res.* **257**, 16–23 (2009).
10. L. Petreanu, D. Huber, A. Sobczyk, K. Svoboda, *Nat. Neurosci.* **10**, 663–668 (2007).
11. J. J. Letzkus, S. B. E. Wolff, A. Lüthi, *Neuron* **88**, 264–276 (2015).
12. G. Major, M. E. Larkum, J. Schiller, *Annu. Rev. Neurosci.* **36**, 1–24 (2013).
13. E. M. Callaway, L. Luo, *J. Neurosci.* **35**, 8979–8985 (2015).
14. T. R. Reardon *et al*., *Neuron* **89**, 711–724 (2016).
15. D. Cai *et al*., *Cereb. Cortex* **29**, 2437–2454 (2019).
16. C. Herry, J. P. Johansen, *Nat. Neurosci.* **17**, 1644–1654 (2014).
17. J. J. Letzkus *et al*., *Nature* **480**, 331–335 (2011).
18. C. Mezzera, G. López-Bendito, *J. Chem. Neuroanat.* **75**, 32–40 (2016).
19. R. Batista-Brito, E. Zagha, J. M. Ratliff, M. Vinck, *Curr. Opin. Neurobiol.* **52**, 172–181 (2018).
20. S. A. Vasquez-Lopez *et al*., *eLife* **6**, e25141 (2017).
21. S. Oláh *et al*., *Nature* **461**, 1278–1281 (2009).
22. M. Capogna, *J. Physiol.* **589**, 1875–1883 (2011).
23. G. Tamás, A. Lörincz, A. Simon, J. Szabadics, *Science* **299**, 1902–1905 (2003).
24. T. Deneux *et al*., *eLife* **8**, e44006 (2019).
25. S. M. Sherman, R. W. Guillery, in *The Synaptic Organization of the Brain*, G. M. Shepherd, Ed. (Oxford Univ. Press, ed. 5, 2004) chap. 8, pp. 311–360.
26. M. M. Halassa, S. Kastner, *Nat. Neurosci.* **20**, 1669–1679 (2017).
27. N. Takahashi, T. G. Oertner, P. Hegemann, M. E. Larkum, *Science* **354**, 1587–1590 (2016).
28. L. A. Anderson, J. F. Linden, *Hear. Res.* **274**, 48–60 (2011).
29. Á. Kulik, S. A. Booker, I. Vida, *Neuropharmacology* **136**, 80–91 (2018).
30. P. R. Roelfsema, A. Holtmaat, *Nat. Rev. Neurosci.* **19**, 166–180 (2018).
31. D. Castaño-Díez, Freezing Analysis, Github (2018); https://doi.org/10.17617/1.8Q.
32. F. Kretschmer, EyeTracker, MPCDF Gitlab (2020); https://doi.org/10.17617/1.8M.
33. F. Kretschmer, PylonRecorder, MPCDF Gitlab (2020); https://doi.org/10.17617/1.8N.
34. F. Kretschmer, D. Unzué, AudioGameGUI, MPCDF Gitlab (2020); https://doi.org/10.17617/1.8O.
35. F. Kretschmer, D. Castaño-Díez, MotionCorrection, MPCDF Gitlab (2020); https://doi.org/10.17617/1.8P.
36. L. B. Naumann, H. Sprekeler, Model code for frequency modulation with presynaptic inhibition, Version 1, Zenodo (2020); https://doi.org/10.5281/zenodo.3886569.

ACKNOWLEDGMENTS

We thank members of the Letzkus and Sprekeler labs, M. S. Fustinana Gueler, J. Gjorgjieva, S. Romano, and J. Keijser for discussions; E. Abs, A. Wrana, S. Junek, F. Vollrath, G. Tushev, and S. Onasch for technical assistance; and L. L. Looger, J. Akerboom, D. S. Kim, the GENIE Project at Janelia Farm, K. Deisseroth, E. S. Boyden, H. Zeng, and K.-K. Conzelmann for generously sharing reagents. **Funding:** Max Planck Society, German Research Foundation (LE 3804/3-1, CRC 1193, and SPP1665 to J.J.L.), European Research Council (StG 335587 to J.J.L.), Human Frontier Science Program (LT000814/2016-L to M.B.P.).
Author contributions: Conceptualization, project administration, writing, and funding acquisition: M.B.P. and J.J.L. Investigation and analyses: M.B.P., J.V., T.D., T.S., and D.-L.P. Computational modeling conceptualization, methodology, and software: L.B.N., F.K., and H.S. Visualization: M.B.P. Supervision: J.J.L. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** All data are available in the main text or the supplementary materials. Code for analysis is available on GitHub or GitLab (*31*–*35*) and code for computational model at Zenodo (*36*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/844/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S14
Table S1
References (*37*–*49*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

14 April 2020; accepted 30 September 2020
10.1126/science.abc2399

SENSORS

Stretchable distributed fiber-optic sensors

Hedan Bai[1]*, Shuo Li[2]*, Jose Barreiros[3], Yaqi Tu[2], Clifford R. Pollock[4], Robert F. Shepherd[1,2,3]†

Silica-based distributed fiber-optic sensor (DFOS) systems have been a powerful tool for sensing strain, pressure, vibration, acceleration, temperature, and humidity in inextensible structures. DFOS systems, however, are incompatible with the large strains associated with soft robotics and stretchable electronics. We develop a sensor composed of parallel assemblies of elastomeric lightguides that incorporate continuum or discrete chromatic patterns. By exploiting a combination of frustrated total internal reflection and absorption, stretchable DFOSs can distinguish and measure the locations, magnitudes, and modes (stretch, bend, or press) of mechanical deformation. We further demonstrate multilocation decoupling and multimodal deformation decoupling through a stretchable DFOS–integrated wireless glove that can reconfigure all types of finger joint movements and external presses simultaneously, with only a single sensor in real time.

Light has primary properties of intensity, wavelength spectrum, polarization, and propagation direction that can be individually detected to reflect disturbances in the propagating medium. On the basis of these properties, sensing through optical fibers has been a useful platform for monitoring mechanical deformation in stiff infrastructures (e.g., bridges, roads, buildings, and geophysics) (*1*–*3*). Distributed fiber-optic sensor (DFOS) systems can be categorized into two approaches: (i) intrinsic, in which a single measurand (e.g., strain or temperature) can be monitored continuously through the length of an optical fiber via Raman, Rayleigh, or Brillouin scattering; or (ii) quasi, such as fiber Bragg grating (FBG) sensors, in which multiple measurands (e.g., strain, pressure, vibration, acceleration, temperature, and humidity) can be monitored with discrete spatial resolution through sensitizing sections of an optical fiber for specific measurements (*1*–*4*). Although commercially available DFOSs have already shown promise for advanced multifunctional sensing in soft materials (*5*, *6*), the inextensible silica optical fibers, bulky laser light source, and expensive detection equipment required for resolving the subtle backscattered wavelength shift become barriers to broad adoption of this technology in research areas such as soft robotics (*7*), stretchable electronics (*8*, *9*), and biomedicine (*10*, *11*).

To apply the abilities of DFOSs to tactile sensation in stretchable optoelectronics, we took inspiration from mammals. Mammals have soft tissues that allow mechanical stimuli to propagate to various depths, where different deformation modalities are detected and processed by distinctive mechanoreceptors (*12*). Thus, mechanically compliant deformation sensors with multiple sensing modalities could give intelligent soft systems the capacity for rich tactile sensations comparable to those of mammals. Myriad classes of intrinsically stretchable sensors have been developed on the basis of resistive (*13*–*15*), capacitive (*16*, *17*) and optical sensing principles (*18*–*22*). These highly stretchable sensors have demonstrated numerous sensing modes for a variety of deformations (e.g., stretch, bend, press, and twist). The sensors, however, share the same limitation: They respond to the different deformation modes with changes of one parameter (resistance, capacitance, optical intensity, or chromaticity). Therefore, unlike mechanoreceptors that distinctively react to certain stimuli, these stretchable sensors cannot easily differentiate deformation modes. To detect complex deformation combinations, multiple sensors are usually geometrically constrained strategically to ensure that each sensor is decoupled and detects one mode of deformation (*18*, *23*). To detect spatial information, a grid-like distribution of sensors is employed in two or three dimensions (*24*–*26*). As sensing becomes more advanced, the sensor system's architectural complexity increases in terms of sensor placement and density. Toward this end, machine learning has recently been adopted in soft sensors that can measure and differentiate between bending and twisting with simple sensor layouts (*27*). Until now, sensing of complex soft-body deformations has been addressed by either complicated system integration or computation methods that involve model training.

We present a multifunctional stretchable optomechanical sensor that we call the stretchable DFOS, inspired by silica-based DFOSs. Stretchable DFOS systems are collateral assemblies of elastomeric lightguides with embedded chromatic patterns. We have designed two chromatic dye patterns as analogies of the two types of conventional DFOSs: (i) a continuum gradient dye pattern, for which we can obtain continuous spatial resolution of single-mode deformation (fig. S1); and (ii) a discrete color-block pattern, for which we can measure and decouple multiple modes of deformations with discrete spatial resolution. To demonstrate the multifunctionality of stretchable DFOSs, our work focused on the design with the discrete color-block pattern, which we name the stretchable lightguide for multimodal sensing (SLIMS). We compare major features of SLIMS, FBG, and intrinsic DFOSs in table S1.

Figure 1, A and B, shows the construction of SLIMS with a lightguide consisting of two polyurethane elastomeric cores—a dyed core doped with absorbing dyes at four discrete locations and a clear core with no dyes—separated by silicone cladding. A white light-emitting diode (LED) is coupled to the dyed core at one end. Two red-green-blue (RGB) sensor chips are placed at the other end, one coupled to the dyed core and one coupled to the clear core (see the "sensor design" and "sensor fabrication" sections in the materials and methods). Our optomechanical sensing approach is based on deformation-induced geometric changes in the optical path of light propagation in SLIMS. Chromatic dyes provide wavelength-selective modulation through light absorption variations; the dual-core structure allows frustrated total internal reflection (TIR). By observing the chromaticity and intensity outputs from the two cores, we can determine the location, magnitude, and modes of deformation (Fig. 1C and movies S1 to S3).

Because the lightguides have cross sections at the millimeter scale, the wave properties of visible light (with wavelengths <700 nm) become unimportant. The white LED fitted to the dyed core input has a viewing angle of 110° so that rays from the same LED also enter the clear core. The doped dyes act as color codes for the spatial information. Because the dye patterns have a depth equal to only a fraction (in this case, ¼) of the dyed core height, very few light rays can pass through the dyes in the undeformed configuration (for a ray diagram, see Fig. 1C, undeformed). Both cores have white outputs when undeformed.

When a dyed region is stretched (ray diagram: see Fig. 1C, stretched), the optical path length in that region increases, inducing more absorption by the dye and a change in output light toward the corresponding color. By contrast, the clear core does not change color owing to insufficient coupling between the two cores at the stretched dyed region. The ray optics simulation (COMSOL Multiphysics, COMSOL Inc.) also confirms this color-changing response (figs. S4 and S5A). Because stretching increases the optical path in both cores, according to the Beer-Lambert law, light outputs for both cores decrease in intensity.

[1]Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY 14853, USA. [2]Department of Materials Science and Engineering, Cornell University, Ithaca, NY 14853, USA. [3]Program of Systems Engineering, Cornell University, Ithaca, NY 14853, USA. [4]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: rfs247@cornell.edu

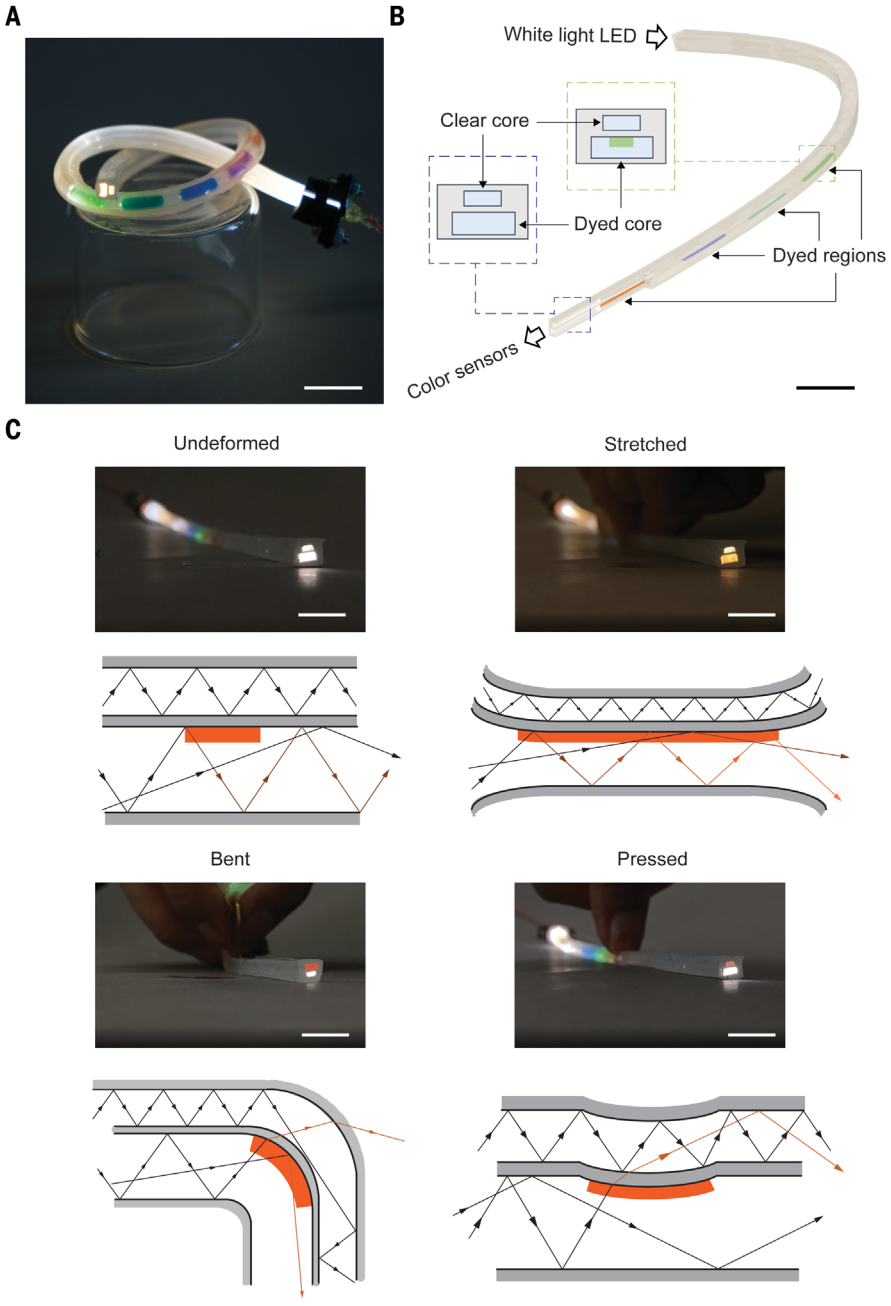


Fig. 1. SLIMS. (**A**) Image of a SLIMS tied into a knot. (**B**) Schematic of the SLIMS showing the discrete dyed regions, the design of the collateral cores, and its coupling to a light source and color sensors. (**C**) Optical outputs and ray diagrams of SLIMS when it is undeformed, stretched, bent, and pressed. Scale bars, 1 cm.

Characterization of SLIMS shows that as any single dyed region is stretched (for details, see the "sensor characterization" section in the materials and methods), no chromaticity change in the clear core is observed (Fig. 2A), whereas chromaticity change with increasing saturation is observed in the dyed core (Fig. 2D). Both cores have output intensities that attenuate linearly in a logarithmic scale (Fig. 2, G and J). On the basis of the dyed core's attenuation, strain sensitivity of SLIMS falls within 2 to 5 dB ε^{-1} (where ε is strain) (see the "sensitivity variation and dye absorbance" section in the supplementary text) over a range of 150% elongation. We obtained 0.1-mm resolution (1% strain) with a 0.5-mm prestrain. We found that prestraining improves the sensor resolution (fig. S6), which is consistent with the stretch characterization results (Fig. 2J). Note the nonlinear intensity response when the yellow-green region (YG) is stretched (Fig. 2G): The YG dye is placed too close to the light source in the spatial transient region of the lightguide (see the "nonlinearity in sensor response" section in the supplementary text).

When a dyed region is in bending configuration (Fig. 1C, bent), most rays in the dyed core can pass through the dye, leading to a change in the output color. Some colored rays that fail to satisfy the critical angle escape the dyed core and enter the clear core. These rays, however, cannot be coupled to the clear core. Thus, the clear core still outputs white light (for a ray-optics simulation, see fig. S5B). With each of the four dyed regions bent to a curvature up to $\kappa = 0.65\ \mathrm{cm}^{-1}$, we measured the chromaticity and intensity responses of both cores (Fig. 2, B, E, H, and K). The chromaticity plots demonstrate that the color response to bending is similar to the stretching response: The clear core output remains white, and the dyed core output changes its color toward the bent dye region (Fig. 2, B and E). Differentiation between stretching and bending can be realized by comparing the intensity output of the clear core, as the intensity is largely unaffected (Fig. 2H) in bending and attenuates substantially in stretching. With the dyed core's attenuation (Fig. 2K), SLIMS' bending sensitivity is calculated to be 7 to 24 $\mathrm{dB{\cdot}cm}^{-1}$. The bending resolution is measured as 0.5° using a digital protractor (fig. S7).

Frustrated TIR takes place when SLIMS is pressed (Fig. 1C, pressed). Rays in the dyed core escape through the dye and couple to the clear core, leading to a colored output in the clear core (Fig. 2C). As the compression force increases, the clear core has more-saturated color output. The dyed core output remains white because there is no increase in optical path length or number of rays that pass through the dye (Fig. 2F). On the basis of the intensity attenuation of the clear core (Fig. 2I), SLIMS has a sensitivity to pressing of 0.9 to 1.2 $\mathrm{dB\ N}^{-1}$ in the range of $2\ \mathrm{N} < F < 5\ \mathrm{N}$ (F, force). This sensitivity can be tuned with the Young's moduli of the constructing elastomers. Compression force resolution is found to be 0.14 N. We observed high repeatability (fig. S8) and close to linear trend (in logarithmic scale) in the intensity attenuation for all dyed regions except YG. This issue is also attributable to the proximity of the YG region to the light source.

We incorporated SLIMS into a 3D printed soft glove, where each finger uses only a single customized SLIMS that allows proprioceptive sensing of three finger joints' movements and exteroceptive sensing of external presses to be captured simultaneously. The proximal, middle, and distal joints of the finger are covered by discrete red, blue, and green dyes, respectively (Fig. 3A; see the "glove fabrication" section in the materials and methods).

For proprioception, we first collected the raw data of RGB intensities from both cores

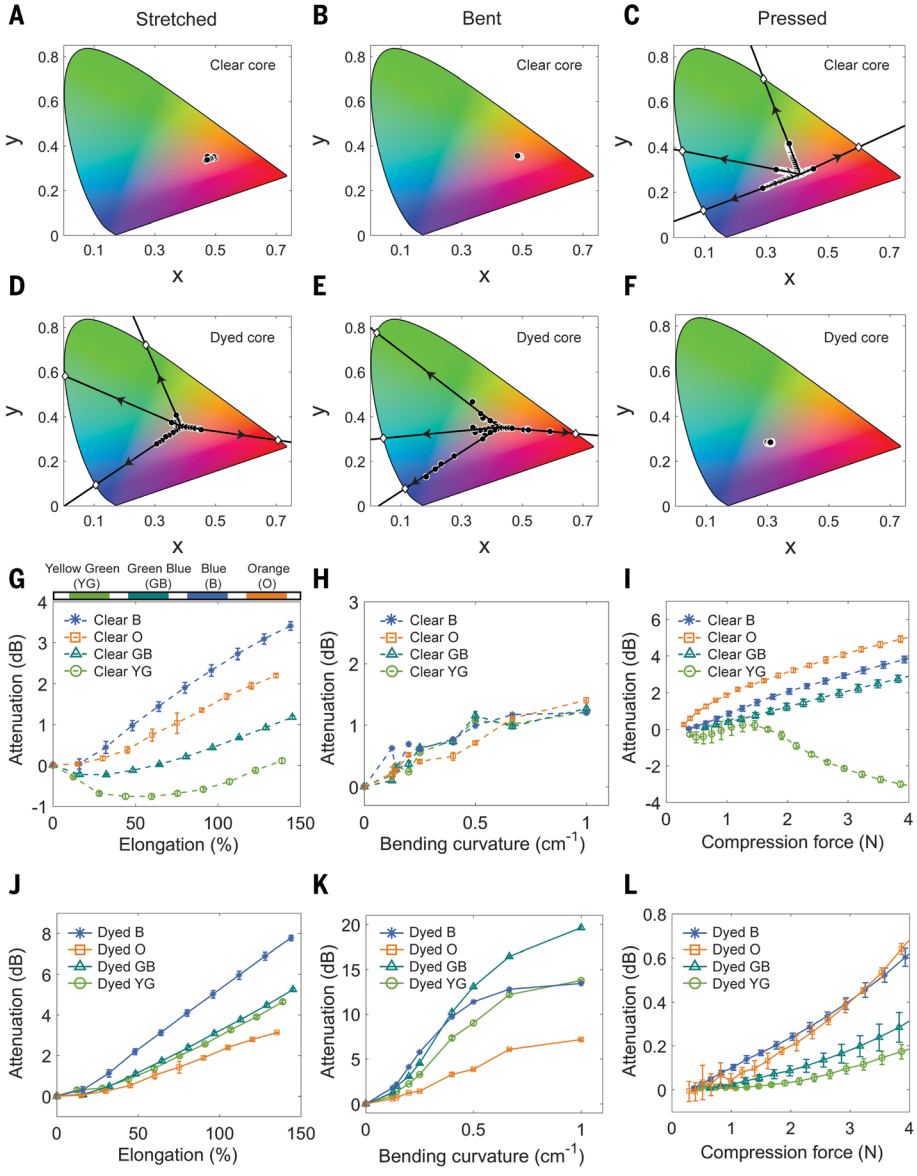


Fig. 2. Characterization of SLIMS in different deformation modes. Chromaticity responses of the clear core when SLIMS is (**A**) stretched, (**B**) bent, or (**C**) pressed in the dyed regions. Chromaticity responses of the dyed core when SLIMS is (**D**) stretched, (**E**) bent, or (**F**) pressed in the dyed regions. Intensity responses of the clear core when SLIMS is (**G**) stretched, (**H**) bent, or (**I**) pressed in the dyed regions. Intensity responses of the dyed core when SLIMS is (**J**) stretched, (**K**) bent, or (**L**) pressed in the dyed regions. Error bars indicate SDs.

as the index finger is bent with different joint angle combinations (Fig. 3B). When there is only one bend from either the proximal or the middle joint (it is difficult to bend only the distal joint), the dyed core immediately changes color from white to either red or blue. We can ascertain both the location and the angle of that bend. When multiple bends coexist, however, we observed a mixture of colors from the bent regions. Therefore, we need to develop a mathematical model to decouple the fusion in both chromaticity and intensity responses (Fig. 3C).

We first measured the dyed core response in SLIMS when only a single joint was bent, and we then repeated this measurement for all three joints. To decouple intensity and chromaticity response, we converted the RGB intensity output into the CIE *xyY* color space (Fig. 3D). As each joint bends from 0° to 90° with an interval of 10°, the normalized intensity output Y (represented by the size of the dots) drops and the chromaticity output xy (represented by the color of the dots) becomes more saturated. Each combined response (represented by the dots) follows a linear trend, producing red, blue, and green single-joint bend reference lines. The space within the pyramid formed by these three reference lines thus includes all possible multijoint bend outputs. More specifically, we can formulate the algorithm to decouple multilocation bending as a vector sum model: Any multijoint bend response [denoted by a random point $P(x, y, Y)$ within the pyramid] can be related to the three single-joint bend reference vectors with different weights, and those weights depict the respective bending angles at each joint (Fig. 3C)

$$\overrightarrow{\mathbf{OP}} = (\text{Proximal Joint Angle}/90°)\,\overrightarrow{\mathbf{OR}} + (\text{Middle Joint Angle}/90°)\,\overrightarrow{\mathbf{OB}} + (\text{Distal Joint Angle}/90°)\,\overrightarrow{\mathbf{OG}}$$

where point O represents the response of the undeformed output with white color and maximum intensity. Converting the intensity attenuation to a logarithmic scale (in units of decibels) yielded vectors $\overrightarrow{\mathbf{OR}}$, $\overrightarrow{\mathbf{OB}}$, and $\overrightarrow{\mathbf{OG}}$ that correspond to the three linearly fitted single-joint bend reference lines. Vector $\overrightarrow{\mathbf{OP}}$, the relative response of an arbitrary multijoint bend configuration, then can be expressed as a vector sum of $\overrightarrow{\mathbf{OR}}$, $\overrightarrow{\mathbf{OB}}$, and $\overrightarrow{\mathbf{OG}}$ with different weights. Because vectors $\overrightarrow{\mathbf{OR}}$, $\overrightarrow{\mathbf{OB}}$, and $\overrightarrow{\mathbf{OG}}$ are linearly independent and span the entire 3D space, this model has unique solutions for the three unknowns: proximal, middle, and distal joint angles. See the supplementary text for model validation. With the vector sum model, we can solve for the three unknown joint angles, given the measured RGB intensities in the dyed core. Figure 3G and movie S4 show that the glove performs decoupling and reconstruction of the three finger joints' movements in real time.

For simultaneous exteroceptive sensing, we achieved the decoupling of external pressing from bending by setting thresholds for the normalized intensity output of the clear core (Fig. 4A and movie S4). In the case of bending only (0 to 10 s), the intensity attenuates slightly; it jumps when the SLIMS is pressed at different joints for multiple times (10 to 30 s). We set this threshold within the range of 0.75 to 1.0. When the normalized intensity exceeds this range, we could determine the presence of a press. To simplify the representation of chromaticity, we extracted the hue value from the RGB intensities of the clear core output (Fig. 4B). The hue representation enabled us to locate the region that is being pressed. The magnitude of the exerted force during pressing also qualitatively scales with this hue value. We were able to reconstruct, in real time, the location and magnitude of an external press when the index finger is bent at multiple locations (Fig. 4C).

The principles of stretchable DFOSs could also be applied to detect measurands beyond those demonstrated in this work and for different applications. For example, movie S5 shows the dynamic color change that could be used for slip detection in robotic grippers.

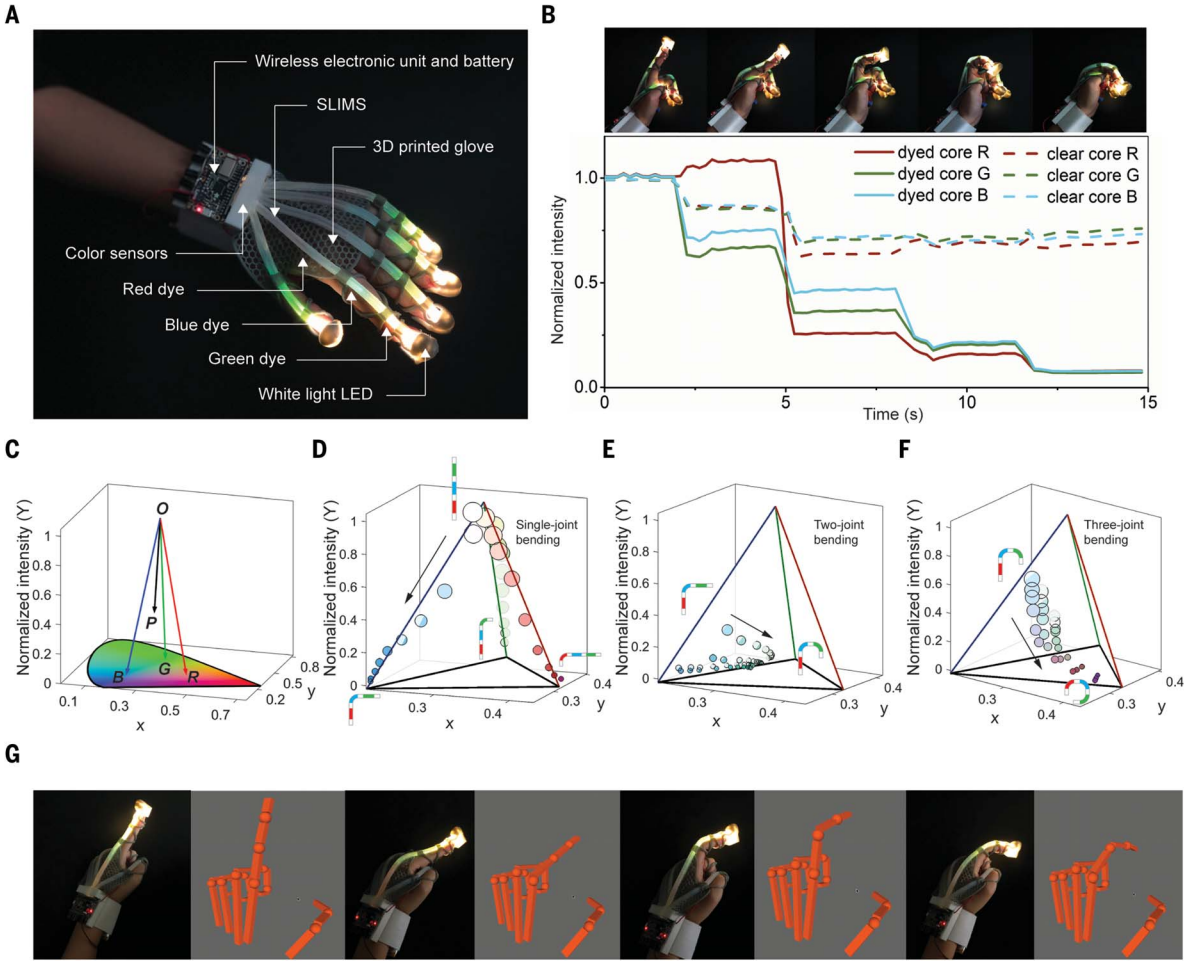


Fig. 3. SLIMS-integrated soft glove and proprioception of multijoint bending. (**A**) Image of the SLIMS-integrated soft glove with LED light source and electronics module. (**B**) Real-time normalized intensity data from both cores under five different multijoint bending configurations. (**C**) Vector sum model for multijoint bend decoupling. (**D**) Single-joint bend measurements. (**E**) Two-joint bend measurements. (**F**) Three-joint bend measurements. (**G**) Real-time reconstruction of multijoint bending using the vector sum model that we derived.

The spatial resolution of the discrete dye pattern is limited by the spacing between the dyes; distributed sensing could be achieved through the gradient dye pattern (fig. S1). We characterized the spatial resolution as 1 mm for the gradient dye in bending configuration (fig. S11; see the "distributed sensing characterization with gradient pattern" section in the supplementary text). Combining the discrete and gradient patterns could enable multilocation detection with continuous spatial resolution (fig. S12).

Our current SLIMS sensor is limited to decoupling three dyed regions for multibend applications, owing to the limited output generated by the RGB color sensor chip (fig. S13A, three color bands). If color sensors with better spectrum resolution (e.g., PixelSensor multispectral sensors, PIXELTEQ; fig. S13B, eight color bands) are used, the number of decouplable dyed regions can be improved to the number of bands that the color sensor can resolve (see the "multi-bend decoupling and color sensor resolution" section in the supplementary text). The cross-sectional dimensions of SLIMS are substantially larger than those of silica optical fibers, owing to the size of the coupled optoelectronic devices; however, the size of SLIMS could be reduced with a tapered design. The theoretical limit in fiber diameter is then comparable to that of commercial silica multimodal fibers (~50-μm diameter; see the "scaling" section in the supplementary text). SLIMS has an attenuation coefficient of 0.68 dB·cm^{-1}, which limits the range of sensing length to tens of centimeters. Multimeter-scale applications can be achieved via a hybrid approach of low-loss optical fibers for transmission and SLIMS for sensing (see the "scaling" section in the supplementary text).

REFERENCES AND NOTES

1. A. D. Kersey, A. Dandridge, *J. Inst. Electron. Radio Eng.* **58**, S99 (1988).
2. N. J. Lindsey, T. C. Dawe, J. B. Ajo-Franklin, *Science* **366**, 1103–1107 (2019).
3. K. O. Hill, G. Meltz, *J. Lit. Technol.* **15**, 1263–1276 (1997).
4. B. Lee, *Opt. Fiber Technol.* **9**, 57–79 (2003).
5. K. C. Galloway *et al.*, *Soft Robot.* **6**, 671–684 (2019).
6. Y.-L. Park, K. Chau, R. J. Black, M. R. Cutkosky, in *Proceedings of the 2007 IEEE International Conference on Robotics and Automation* (IEEE, 2007), pp. 1510–1516.
7. S. Li, H. Bai, R. F. Shepherd, H. Zhao, *Angew. Chem. Int. Ed.* **58**, 11182–11204 (2019).
8. S. Niu *et al.*, *Nat. Electron.* **2**, 361–368 (2019).
9. K. Takei, W. Gao, C. Wang, A. Javey, *Proc. IEEE* **107**, 2155–2167 (2019).
10. H. U. Chung *et al.*, *Science* **363**, eaau0780 (2019).
11. S.-H. Byun *et al.*, *Sci. Adv.* **5**, eaay0418 (2019).
12. V. E. Abraira, D. D. Ginty, *Neuron* **79**, 618–639 (2013).
13. T. Yamada *et al.*, *Nat. Nanotechnol.* **6**, 296–301 (2011).
14. M. D. Dickey, *Adv. Mater.* **29**, 1606425 (2017).
15. C. Keplinger *et al.*, *Science* **341**, 984–987 (2013).
16. A. Frutiger *et al.*, *Adv. Mater.* **27**, 2440–2446 (2015).
17. S. Li, B. N. Peele, C. M. Larson, H. Zhao, R. F. Shepherd, *Adv. Mater.* **28**, 9770–9775 (2016).
18. H. Zhao, K. O'Brien, S. Li, R. F. Shepherd, *Sci. Robot.* **1**, eaai7529 (2016).
19. M. Kolle *et al.*, *Adv. Mater.* **25**, 2239–2245 (2013).
20. J. Guo *et al.*, *Adv. Mater.* **28**, 10244–10249 (2016).
21. A. Leber, B. Cholst, J. Sandt, N. Vogel, M. Kolle, *Adv. Funct. Mater.* **29**, 1802629 (2018).
22. R. Ghaffari *et al.*, *Adv. Funct. Mater.* **30**, 1907269 (2019).
23. R. L. Truby *et al.*, *Adv. Mater.* **30**, e1706383 (2018).
24. M. Ramuz, B. C.-K. Tee, J. B.-H. Tok, Z. Bao, *Adv. Mater.* **24**, 3223–3227 (2012).

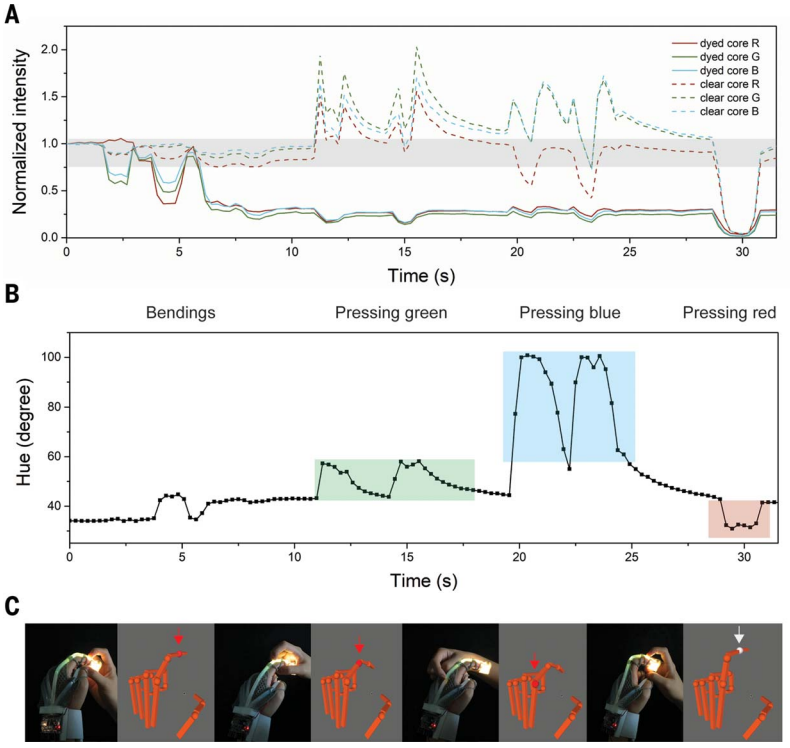


Fig. 4. SLIMS-integrated soft-glove decoupling combined proprioception and exteroception. (**A**) Real-time normalized RGB intensities of both cores under bending and pressing on different joints from 0 to 30 s. The threshold for press detection is denoted by the gray box. (**B**) Real-time hue values derived from the normalized RGB intensities under the same deformations from 0 to 30 s. The colored boxes show hue values that determine the pressed joints. (**C**) Real-time reconstruction of combined proprioception and exteroception.

25. A. Levi, M. Piovanelli, S. Furlan, B. Mazzolai, L. Beccai, *Sensors* **13**, 6578–6604 (2013).
26. P. A. Xu *et al.*, *Sci. Robot.* **4**, eaaw6304 (2019).
27. I. M. Van Meerbeek, C. M. De Sa, R. F. Shepherd, *Sci. Robot.* **3**, eaau2489 (2018).

ACKNOWLEDGMENTS

We thank K. Wang for conducting tensile tests and A. Cornwell for helping with circuit soldering. **Funding:** This work was supported by the NSF EFRI program (grant EFMA-1830924), the Air Force Office of Scientific Research (grant FA9550-18-1-0243), Cornell Technology Acceleration and Maturation (CTAM), National Institute of Food and Agriculture (grant 2019-67021-29225), and the Office of Naval Research (grant N00014-17-1-2837). Part of the study was performed at the Cornell NanoScale Facility, a member of the National Nanotechnology Coordinated Infrastructure, which was supported by the NSF (grant NNCI-1542081), and at the Cornell Center for Materials Research Shared Facilities, which was supported through the NSF MRSEC program (grant DMR-1719875). **Author contributions:** Concept, design, and study direction: H.B. and R.F.S.; device fabrication: H.B., S.L., and J.B.; experimental validation: H.B., S.L., Y.T., and R.F.S.; data analysis: H.B., S.L., and R.F.S.; theoretical modeling: H.B., C.R.P., and R.F.S..; manuscript writing: H.B., S.L., and R.F.S.; manuscript editing: H.B., S.L., C.R.P., and R.F.S.; supervision: C.R.P. and R.F.S.; project administration: R.F.S. **Competing interests:** The stretchable DFOS systems presented in this work have been filed under an international patent application (no. 62/642,407) for "Waveguide and Sensor Based on Same." The listed inventors are R.F.S., H.B., S.L., and Y.T. The intellectual property is being licensed by Organic Robotics Corporation. R.F.S. is the founder of Organic Robotics Corporation. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the main text and the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/848/suppl/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S20
Tables S1 and S2
References (*28*–*31*)
Movies S1 to S5
Code S1

12 December 2019; accepted 5 October 2020
10.1126/science.aba5504

TRANSLATION REGULATION

A phosphorylation-regulated eIF3d translation switch mediates cellular adaptation to metabolic stress

Adam M. Lamper*†, Rebecca H. Fleming*, Kayla M. Ladd*, Amy S. Y. Lee‡§

Shutoff of global protein synthesis is a conserved response to cellular stresses. This general phenomenon is accompanied by the induction of distinct gene programs tailored to each stress. Although the mechanisms driving repression of general protein synthesis are well characterized, how cells reprogram the translation machinery for selective gene expression remains poorly understood. Here, we found that the noncanonical 5′ cap-binding protein eIF3d was activated in response to metabolic stress in human cells. Activation required reduced CK2-mediated phosphorylation near the eIF3d cap-binding pocket. eIF3d controls a gene program enriched in factors important for glucose homeostasis, including members of the mammalian target of rapamycin (mTOR) pathway. eIF3d-directed translation adaptation was essential for cell survival during chronic glucose deprivation. Thus, this mechanism of translation reprogramming regulates the cellular response to metabolic stress.

During stress, cells undergo substantial alterations to protein synthesis. Global translation is inhibited, whereas the translation of a subset of mRNAs encoding stress-responsive proteins is considerably up-regulated. Translation shutoff is mediated through inactivation of core translation initiation factors, including the 5′ cap-binding protein eukaryotic translation initiation factor 4E (eIF4E) (*1*). How initiation factor activity is modified during stress to allow some mRNAs to escape shutoff despite carrying a conserved 5′ cap structure remains unknown. The eIF3d subunit of the 13-subunit eukaryotic initiation factor complex 3 (eIF3) is a cap-binding protein that directs noncanonical translation of the transcription factor *Jun* (*2*), raising the possibility that eIF3d could be harnessed for stress-induced translation.

To test this hypothesis, we examined changes to eIF3d-mediated translation of the *Jun* mRNA during metabolic stress in human embryonic kidney (HEK) 293T cells. Conditions of nutrient depletion caused a marked decrease in general protein synthesis (fig. S1A), yet chronic glucose depletion led to a pronounced increase in Jun protein levels (Fig. 1A). The increase in Jun expression was stress specific and did not occur with glutamine or serum starvation (Fig. 1A). Analysis of *Jun* mRNA distribution in polysome fractions revealed that elevation of Jun protein synthesis during glucose deprivation was due to increased ribosome association and translation efficiency (Fig. 1, B and C). By contrast, the eIF4E-dependent control mRNA *PSMB6* shifted to monosome and free RNA fractions under the same glucose deprivation conditions (Fig. 1, B and C). Thus, the ability to engage eIF3d is likely key to enabling the *Jun* mRNA to escape translation shutoff. We next investigated whether nutrient starvation directly changed eIF3d cap-binding activity. An autoinhibitory conformation restricts eIF3d and necessitates that the subunit must function as part of the 13-subunit eIF3 complex for 5′ cap recognition (*2*). Taking advantage of a natural HIV-1 protease (PR) cleavage site in eIF3d that occurs immediately before the cap-binding domain (fig. S1B) (*3*), we used immunoprecipitation and on-bead cleavage to isolate eIF3d cap-binding activity separately from other eIF3 complex functions (*4*–*6*) (fig. S1, C to G). eIF3d cap binding to *Jun* mRNA increased 10-fold specifically during glucose deprivation (Fig. 1D). Thus, a cellular mechanism can reprogram eIF3d cap-binding activity and translation regulation during nutrient depletion.

Protein phosphorylation is a major mechanism regulating translation factor function (*1*). Using [^{32}P]-orthophosphate labeling, we found that 48 hours after glucose starvation, the eIF3 complex underwent a phosphorylation switch, with loss of modification of an ~70-kDa protein (Fig. 1E). Protein complex digestion with HIV-1 PR confirmed that the phosphorylated protein was eIF3d (Fig. 2A). Mass spectrometry mapped the phosphorylation sites in eIF3d to positions S528 and S529 (Fig. 2B and fig. S2A).

Department of Biology, Brandeis University, Waltham, MA 02453, USA.
*These authors contributed equally to this work.
†Present address: Department of Biological Sciences, Columbia University, New York, NY 10027, USA.
‡Corresponding author. Email: amysiyinglee@gmail.com
§Present address: Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA, and Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.

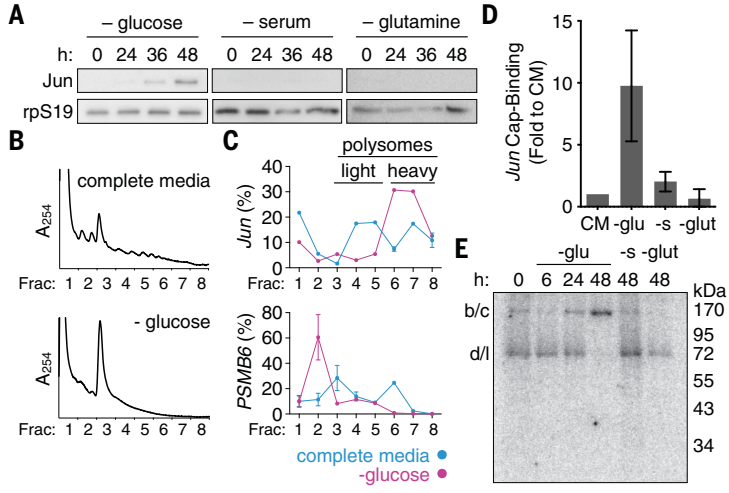


Fig. 1. eIF3d-specialized translation increases during chronic glucose deprivation. (**A**) Jun and rpS19 protein levels in HEK293T cells after nutrient deprivation. (**B** and **C**) Polysome association of *Jun* or *PSMB6* mRNA in response to glucose deprivation in HEK293T cells. *PSMB6* is a control for eIF4E-dependent translation. mRNA abundance is expressed as a percentage of total transcript recovered from all fractions. Results are plotted as the mean ± SD from a representative quantitative reverse transcription polymerase chain reaction (qRT-PCR) performed in duplicate. The results of (A) to (C) are representative of three biological replicates. (**D**) eIF3d binding to the 5′ cap of *Jun* mRNA in HEK293T cells upon nutrient deprivation. eIF3d cap binding is quantified as levels of *Jun* transcripts determined by reverse transcription and quantitative PCR, in HIV-1 PR-treated eIF3d immunoprecipitation samples compared with total input *Jun* RNA. Cap binding is normalized to samples prepared from cells grown in complete media (CM). glu, glucose; s, serum; glut, glutamine. Results are shown as the mean ± SD of three independent experiments. (**E**) Phosphorylation of eIF3 subunits after nutrient deprivation. Shown is a phosphor image of SDS–polyacrylamide gel electrophoresis (PAGE) gel resolving eIF3 subunit phosphorylation levels as detected by [^{32}P]-orthophosphate labeling of HEK293T cells and immunoprecipitation of the eIF3 complex. The results are representative of two independent experiments.

Fig. 2. CK2 phosphorylation inhibits eIF3d cap-binding activity. (**A**) Phosphorylated eIF3 subunit identification by selective cleavage with HIV-1 PR. HIV-1 PR was incubated with eIF3 complex immunoprecipitated from [^{32}P]-orthophosphate-labeled HEK293T cells. HIV-1 PR cleavage of eIF3d leads to faster gel migration. (**B**) Location of phosphorylated eIF3d residues S528 and S529 (blue) identified by mass spectrometry. (**C**) Validation of eIF3d phosphorylation sites using eIF3d-HA mutants by [^{32}P]-orthophosphate labeling in HEK293T cells. Coomassie staining is shown as a loading control. WT, wild-type. (**D** and **E**) Fold change in eIF3d binding to the 5′ cap of *Jun* mRNA by phosphoinhibitory (S528/529N) or phosphomimetic (S528/529D) mutants in HEK293T cells. Cap-binding activity was measured as in Fig. 1D and normalized to WT eIF3d. Results are shown as the mean ± SD of three independent experiments. (**F**) Association of *Jun* transcripts with ribosomal complexes in HEK293T cells in which endogenous eIF3d was replaced with eIF3d mutants through coexpression of shRNA-resistant mutants and an eIF3d-targeting shRNA. The results are representative of two independent experiments, and ribosome association was determined as in Fig. 1C. (**G**) In vitro phosphorylation of the recombinant 12-subunit eIF3 complex by incubation with purified CK2 and [γ^{32}P] ATP. Autophosphorylated CK2 subunits are indicated. (**H**) CK2 inhibition in HEK293T cells with 10 μM CX-4945 or CX-5011 decreases eIF3d phosphorylation levels as assessed by [^{32}P]-orthophosphate labeling. The results in (A), (C), (G), and (H) are representative of three independent experiments.

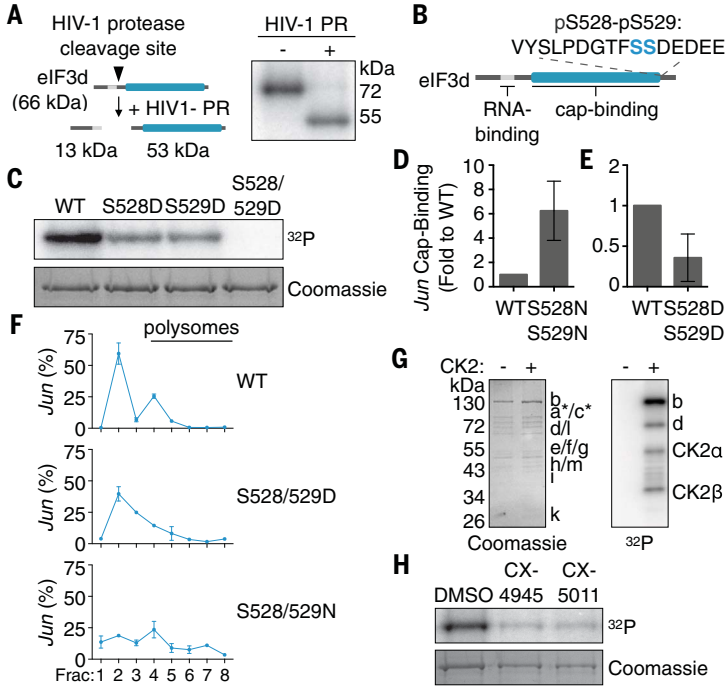
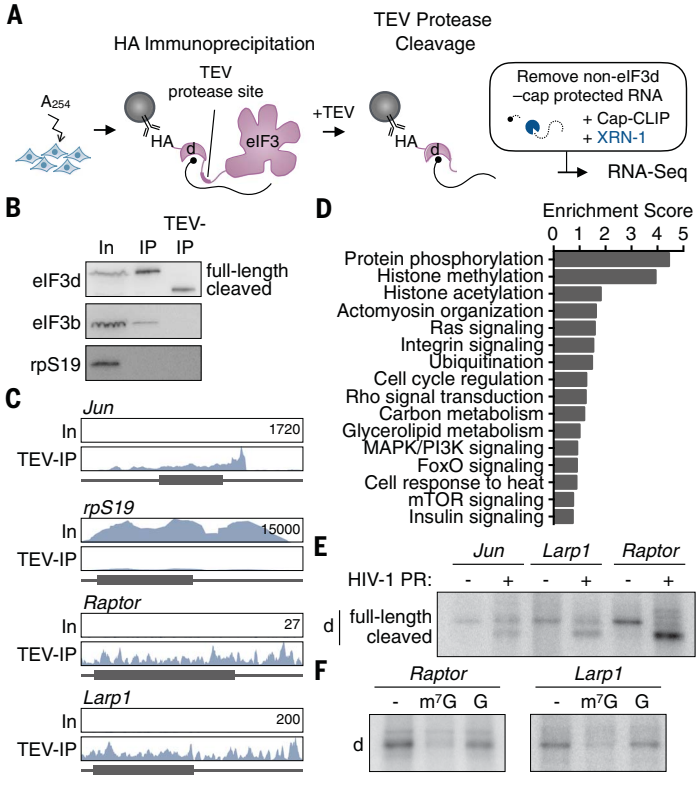


Fig. 3. eIF3d^{TEV} Subunit-Seq identifies cap-binding targets enriched in regulators of cell metabolism. (**A**) Schematic of eIF3d^{TEV} Subunit-Seq methodology. Position R131 within eIF3d was selected for TEV site insertion to minimize perturbation of eIF3d function because this linker region exhibits high phylogenetic variation and is naturally targeted by HIV-1 PR (*3*) (fig. S1B). See text for details. (**B**) Immunoblot of eIF3d^{TEV} Subunit-Seq samples. In, Input lysates; IP, anti-HA eIF3d immunoprecipitants; TEV-IP, IP samples after on-bead incubation with TEV protease. Results are representative of three independent experiments. (**C**) Read mapping to respective transcripts from eIF3d^{TEV} Subunit-Seq performed in glucose-deprived cells. The schematics represent the transcript architecture: Thin lines are the 5′ and 3′ UTRs and thick lines are the open reading frames. The annotated *y*-axis maximum is set equivalently for within each transcript for input (In) and TEV-IP samples. (**D**) Gene ontology analysis of eIF3d cap-binding targets from glucose-deprived cells. (**E**) Phosphor image of SDS-PAGE gel resolving RNase-protected ^{32}P-cap-labeled 5′ UTRs. Recombinant eIF3 complex was cross-linked to target RNAs in vitro and treated with RNase. Cap protection by eIF3d leads to the formation of a radiolabeled covalent eIF3d-cap complex. HIV-1 PR cleavage was performed to validate cross-linked subunit identity as eIF3d, as shown through the shifted migration of the cap-cross-linked complex. (**F**) In vitro eIF3d cap-cross-linking to *Raptor* and *Larp1* 5′ UTRs specifically requires a methylated cap structure. The competitor methylated cap analog m^{7}GpppG (m^{7}G) blocks eIF3d cross-linking to the 5′ UTR, unlike the unmethylated cap analog GpppG (G). The results in (E) and (F) are representative of three independent experiments.



Phosphorylation was reduced with single-site mutations and completely abolished in an eIF3d S528/529D double mutant (Fig. 2C). The eIF3d phosphorylation sites are in a location capable of influencing 5′ cap recognition, proximal to the cap-binding domain (*2*) (Fig. 2B and fig. S2B). In support of a regulatory role of phosphorylation, phosphoinhibitory (S528N/S529N) mutations increased eIF3d 5′ cap-binding activity in cells (Fig. 2D and S2C). Conversely, phosphomimetic (S528D/S529D) eIF3d mutations inhibited 5′ cap-dependent *Jun* mRNA interactions (Fig. 2E) and mirrored the reduced ability of eIF3d to interact with *Jun* when cells were in nutrient-replete conditions (Fig. 1, D and E). We next measured *Jun* translation efficiency in cells engineered to express eIF3d mutants exclusively through stable knockdown and rescue with short hairpin (shRNA)–resistant plasmids (fig. S2, D and E). *Jun* mRNA association was lost in the polysome fractions of cells expressing eIF3d S528D/S529D and gained in cells expressing eIF3d S528N/S529N

(Fig. 2F and fig. S2F). Thus, phosphorylation under nutrient-replete conditions represses eIF3d cap binding, and loss of phosphorylation upon chronic glucose deprivation relieves this inhibition.

The eIF3d phosphorylation sites are each predicted substrates of the protein kinase CK2, which recognizes the sequence motif S*/T*XXE/D with an enrichment in downstream acidic residues (fig. S3A) (*7*, *8*). CK2 is regulated by environmental stimuli (*9*), and CK2-substrate complex formation is inhibited in low-glucose conditions (*10*), in agreement with nutrient-dependent phosphorylation of eIF3d (Fig. 1E). We confirmed that CK2 was capable of modifying eIF3d in vitro in the context of the full eIF3 complex (Fig. 2G). Treatment of cells with the CK2 active-site inhibitors CX-4945 or CX-5011 (*11*) led to a loss of eIF3d phosphorylation (Fig. 2H). By contrast, eIF3d phosphorylation was unaffected by treatment with staurosporine (fig. S3B), a broadly acting kinase inhibitor for which CK2 is atypically insensitive (*12*). Additionally, CK2 inhibition increased the ability of wild-type, but not phosphomimetic, eIF3d to bind the 5′ cap (fig. S3C), further demonstrating a role for the CK2-dependent phosphorylation in regulating eIF3d.

Activation of eIF3d cap-binding activity during glucose deprivation suggests that an eIF3d-controlled gene program may be critical to supporting the cellular response to this metabolic perturbation. However, because eIF3d is an essential gene and 5′ cap recognition occurs as part of the ~800-kDa 13-subunit eIF3 complex (*2*, *5*), technical restrictions have prevented a comprehensive identification of eIF3d target mRNAs (*2*, *13*, *14*). To overcome this challenge, we developed Subunit-Seq, a method that allows for the specific release and isolation of an individual subunit domain and its associated mRNAs away from a larger multisubunit complex. We created a cell line expressing an engineered eIF3d protein with an internal tobacco etch virus (TEV) protease cleavage site proximal to the cap-binding domain and a C-terminal hemagglutinin (HA) affinity tag (Fig. 3A and fig. S4A). To perform Subunit-Seq, we isolated HA-tagged eIF3 complexes and used TEV protease to induce on-bead cleavage and specific isolation of the eIF3d cap-binding domain–RNA complexes (Fig. 3, A and B). Contaminating RNAs that did not have eIF3d-protected 5′ cap structures were removed with on-bead pyrophosphatase decapping and 5′ exoribonuclease digestion. This methodology can be readily adapted to separate the functions of other multisubunit RNA-binding protein complexes through rational insertion of TEV protease sites.

eIF3d^{TEV} Subunit-Seq revealed that eIF3d binds the 5′ cap of 668 transcripts in glucose-deprived HEK293T cells (Fig. 3C; fig. S4, B and C; and table S1). In agreement with a starvation-induced increase in cap binding, ~70% of the eIF3d targets exhibited a two-fold or higher increase in eIF3d binding upon glucose deprivation (fig. S4, D to F, and tables S2 and S3). Gene ontology analysis of the eIF3d targets identified an enrichment in cell metabolism and glucose homeostasis pathways, including mammalian target of rapamycin (mTOR), mitogen-activated protein kinase (MAPK)/phosphatidylinositol 3-kinase (PI3K), and forkhead box O (FoxO) signaling (Fig. 3D). Indeed, eIF3d targets included *Regulatory-associated protein of mTOR* (*Raptor*), an essential component of the mTOR complex 1 (mTORC1) (*15*), and *La-related protein 1* (*Larp1*), a protein that blocks translation of 5′ terminal oligopyrimidine motif-containing RNAs during mTOR inhibition (Fig. 3C) (*16*, *17*). We biochemically validated targets using purified recombinant eIF3 complex and in vitro transcribed mRNAs (*2*) and observed that eIF3d bound to and protected the ^{32}P-labeled cap structure of *Raptor* and *Larp1* 5′ untranslated regions (UTRs) from RNase digestion in vitro (Fig. 3E and fig. S4G). This interaction required a 5′ methylated cap structure, with binding competed away by the competitor ligand m^{7}GpppG (Fig. 3F). Using selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) RNA structure analysis, we identified a stem loop structure in the *Raptor* 5′ UTR that is conserved with the characterized eIF3–*Jun* RNA-binding determinant (*4*), including a U-rich loop, an upper C-rich internal bulge, and a multihelical junction (fig. S5, A and B, and fig. S6). Disruption of this *Raptor* 5′ UTR stem loop by deletion or mutagenesis was sufficient to block eIF3d cap binding (fig. S5C) and translation (fig. S5, D to F), suggesting that the presence of conserved secondary structure may allow coordinated regulation of the eIF3d-directed gene program.

We next investigated whether up-regulation of the eIF3d-directed gene program through the phosphorylation switch provides an adaptive response to glucose deprivation. Lack of glucose triggered increased cell death, rounded cell morphology, and poor attachment in cells expressing a phosphomimetic eIF3d mutant (Fig. 4, A and B, and fig. S7A). By contrast, CK2 inhibition enhanced the cell survival phenotype in cells expressing wild-type, but

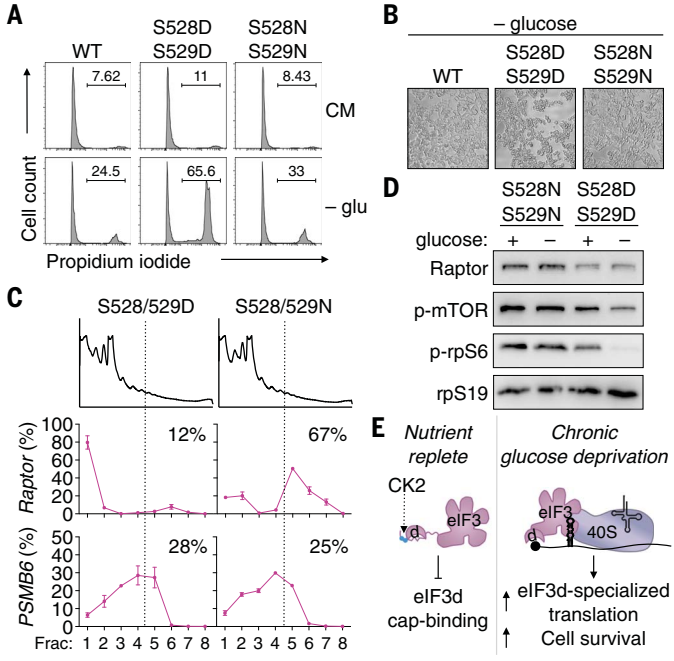


Fig. 4. Regulation of eIF3d-specialized translation through a phosphorylation switch promotes cell survival during chronic glucose deprivation. (**A**) Analysis of apoptotic cell percentage by propidium iodide staining and flow cytometry in eIF3d phosphorylation mutant cell lines. Results in (A) are representative of two independent experiments. (**B**) Representative images of the cell morphology after 48 hours of glucose deprivation in eIF3d phosphorylation mutant cell lines. Results in (B) are representative of three independent experiments. (**C**) Association of *Raptor* mRNA with ribosomal complexes in eIF3d mutant cell lines during glucose deprivation. The percentage of total RNA associated with heavy polysomes (i.e., more than three ribosomes) is indicated. Results are representative of two independent experiments, and ribosome association was determined as in Fig. 1C. (**D**) Immunoblot of mTORC1 pathway activation in eIF3d phosphorylation mutant cell lines during glucose deprivation. Results in (D) are representative of three independent experiments. (**E**) Model for phosphorylation-mediated regulation of eIF3d-specialized translation during metabolic stress.

not phosphomimetic, eIF3d (fig. S7B). Furthermore, dephosphorylated eIF3d was essential for the target mRNA *Raptor* to bypass starvation-induced global shutoff (Fig. 4C and fig. S7, C and D), whereas phosphomimetic eIF3d-expressing cells exhibited decreased Raptor protein levels (Fig. 4D). Raptor is critical for substrate recruitment and mTORC1 kinase activity (*15*). Correspondingly, during glucose deprivation, incorrect control of eIF3d phosphorylation status led to decreased levels of activated mTOR and reduced phosphorylation of rpS6, a downstream target of the mTOR pathway (*15*, *18*) (Fig. 4D). Although mTOR is rapidly inactivated within a few hours of nutrient deprivation, long-term starvation reactivates it and promotes cell survival (*19*, *20*). A protective response during chronic endoplasmic reticulum stress is also dependent on partial rescue of mTOR activity (*13*, *21*). This response is blocked by shRNA-mediated knockdown of eIF3d (*13*), which could be due to the inability of *Raptor* and the eIF3d-mediated gene program to be translated (fig. S7, E and F, and table S1).

Our results establish a model in which a phosphorylation-regulated switch in eIF3d cap-binding activity enables cellular control of eIF3d-specialized translation in response to metabolic state (Fig. 4E). These findings uncover a parallel between stress-induced regulation of eIF3d and the essential canonical translation factors eIF2 and eIF4E (*1*). During glucose deprivation, phosphorylation-regulated inhibition of eIF2 and eIF4E, through hypophosphorylation and activation of the 4E-BPs, allows cells to block general protein synthesis. The conserved but opposing regulation of eIF4E and eIF3d in response to glucose deprivation reveals that the balance of canonical versus noncanonical cap-dependent translation is a key decision point for cellular adaptation to metabolic state.

REFERENCES AND NOTES

1. B. Liu, S. B. Qian, *Wiley Interdiscip. Rev. RNA* **5**, 301–315 (2014).
2. A. S. Lee, P. J. Kranzusch, J. A. Doudna, J. H. Cate, *Nature* **536**, 96–99 (2016).
3. S. Jäger *et al*., *Nature* **481**, 365–370 (2011).
4. A. S. Lee, P. J. Kranzusch, J. H. Cate, *Nature* **522**, 111–114 (2015).
5. K. Asano *et al*., *J. Biol. Chem.* **272**, 27042–27052 (1997).
6. Y. Hashem *et al*., *Nature* **503**, 539–543 (2013).
7. F. Meggio, L. A. Pinna, *FASEB J.* **17**, 349–368 (2003).
8. N. Blom, T. Sicheritz-Pontén, R. Gupta, S. Gammeltoft, S. Brunak, *Proteomics* **4**, 1633–1649 (2004).
9. O. Filhol, C. Cochet, *Cell. Mol. Life Sci.* **66**, 1830–1839 (2009).
10. S. Welker *et al*., *Int. J. Biochem. Cell Biol.* **45**, 2786–2795 (2013).
11. R. Battistutta *et al*., *Biochemistry* **50**, 8478–8488 (2011).
12. F. Meggio *et al*., *Eur. J. Biochem.* **234**, 317–322 (1995).
13. B. J. Guan *et al*., *Mol. Cell* **68**, 885–900.e6 (2017).
14. C. de la Parra *et al*., *Nat. Commun.* **9**, 3068 (2018).
15. D. H. Kim *et al*., *Cell* **110**, 163–175 (2002).
16. S. Hong *et al*., *eLife* **6**, e25237 (2017).
17. R. M. Lahr *et al*., *eLife* **6**, e24146 (2017).
18. G. G. Chiang, R. T. Abraham, *J. Biol. Chem.* **280**, 25485–25490 (2005).
19. L. Yu *et al*., *Nature* **465**, 942–946 (2010).
20. K. J. Condon, D. M. Sabatini, *J. Cell Sci.* **132**, jcs222570 (2019).
21. B. J. Guan *et al*., *J. Biol. Chem.* **289**, 12593–12611 (2014).

ACKNOWLEDGMENTS

We thank S. T. Aoki, K. Chat, P. Garrity, P. J. Kranzusch, J. K. Nuñez, and members of the Lee laboratory for discussions and S. R. Hildebrand, A. D. Yu, W. Xu, and R. Tomaino for experimental advice. **Funding:** This work was funded by the Charles H. Hood Foundation, the Searle Scholars Program, the Pew Biomedical Scholars Program, and a Sloan Research Fellowship (A.S.Y.L). R.H.F. was supported by NIH T32 GM007122.
Author contributions: The project was conceived by A.S.Y.L, and experiments were performed by A.M.L., R.H.F., K.M.L., and A.S.Y.L. The manuscript was written by A.S.Y.L with editing from A.M.L., R.H.F., and K.M.L., and all authors support the conclusions.
Competing interests: The authors declare no competing interests.
Data and materials availability: All sequencing data have been deposited in the Gene Expression Omnibus under accession number GSE158249.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/853/suppl/DC1
Materials and Methods
Figs. S1 to S7
Tables S1 to S3
References (*22*–*42*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

29 January 2020; resubmitted 20 August 2020
Accepted 24 September 2020
10.1126/science.abb0993

Corrected 12 November 2020. See full text.

CORONAVIRUS

Neuropilin-1 facilitates SARS-CoV-2 cell entry and infectivity

Ludovico Cantuti-Castelvetri[1,2]*, Ravi Ojha[3]*, Liliana D. Pedro[1,2]*, Minou Djannatian[1,2]*, Jonas Franz[4,5,6]*, Suvi Kuivanen[7]*, Franziska van der Meer[4], Katri Kallio[3], Tuğberk Kaya[1,2,8], Maria Anastasina[3,9], Teemu Smura[7], Lev Levanov[7], Leonora Szirovicza[7], Allan Tobi[10], Hannimari Kallio-Kokko[11], Pamela Österlund[12], Merja Joensuu[13], Frédéric A. Meunier[13], Sarah J. Butcher[3,9], Martin Sebastian Winkler[14], Brit Mollenhauer[15,16], Ari Helenius[17], Ozgun Gokce[8], Tambet Teesalu[3,19,20], Jussi Hepojoki[5,21], Olli Vapalahti[7,11,22], Christine Stadelmann[4], Giuseppe Balistreri[3,18]†, Mikael Simons[1,2,23]†

The causative agent of coronavirus disease 2019 (COVID-19) is the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). For many viruses, tissue tropism is determined by the availability of virus receptors and entry cofactors on the surface of host cells. In this study, we found that neuropilin-1 (NRP1), known to bind furin-cleaved substrates, significantly potentiates SARS-CoV-2 infectivity, an effect blocked by a monoclonal blocking antibody against NRP1. A SARS-CoV-2 mutant with an altered furin cleavage site did not depend on NRP1 for infectivity. Pathological analysis of olfactory epithelium obtained from human COVID-19 autopsies revealed that SARS-CoV-2 infected NRP1-positive cells facing the nasal cavity. Our data provide insight into SARS-CoV-2 cell infectivity and define a potential target for antiviral intervention.

An outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections has caused a pandemic associated with a severe acute pulmonary disease named COVID-19 (coronavirus disease 2019) (*1*). A related coronavirus, SARS-CoV, led to a much smaller outbreak in 2003, possibly due to infection occurring predominantly in the lower respiratory system, whereas SARS-CoV-2 spreads rapidly through active pharyngeal viral shedding (*2*). Despite these differences, uptake of both viruses is mediated by the same cellular receptor: angiotensin-converting enzyme 2 (ACE2) (*3*–*5*). One hypothesis to explain the enhanced spreading of SARS-CoV-2 is the presence of a polybasic furin-type cleavage site, RRAR^S, at the S1-S2 junction in the SARS-CoV-2 spike (S) protein that is absent in SARS-CoV (*6*). Similar sequences are found in the S proteins of many other pathogenic human viruses, including Ebola, HIV-1, and highly virulent strains of avian influenza (*6*, *7*). The presence of the polybasic cleavage site in SARS-CoV-2 results in enhanced pathogenicity by priming the fusion activity (*8*) and could potentially create additional cell surface receptor binding sites. Proteolytic cleavage of RRAR^S by furin exposes a conserved C-terminal motif, $RXXR_{OH}$ [where R is arginine and X is any amino acid; R can be substituted by lysine (K)], in the S protein. Such C-terminal sequences that conform to the C-end rule (CendR) are known to bind to and activate neuropilin (NRP1 and NRP2) receptors at the cell surface (*9*). Recent cryo–electron microscopy structures of the SARS-CoV-2 S protein demonstrated that the S1-S2 junction is part of a solvent-exposed loop and is therefore accessible for receptor interactions (*10*, *11*).

To determine whether SARS-CoV-2 can use NRP1 for virus entry and infectivity, we generated lentiviral particles pseudotyped with the SARS-CoV-2 S protein. Pseudoviruses are well suited for virus entry assays, as they allow viral entry to be distinguished from other virus life-cycle steps. Human embryonic kidney 293 T (HEK-293T) cells, which have almost no detectable *ACE2* and *NRP1* transcripts (fig. S1), were transfected with plasmids that encode the two established host factors (*4*), human ACE2 and the transmembrane protease serine 2 (TMPRSS2), or NRP1. When expressed alone, ACE2 rendered cells susceptible to infection (Fig. 1A). Although NRP1 did not promote infection in HEK-293T cells, its coexpression with ACE2 and TMPRSS2 markedly enhanced infection (Fig. 1, A and B). NRP1 expression increased infection in Caco-2 cells, which endogenously express ACE2 (*12*) (Fig. 1C and fig. S1D), showing that NRP1 can potentiate infection in the presence of other host factors. To test the specificity of NRP1-dependent virus entry, we developed monoclonal antibodies (mAbs) that were designed to functionally block the extracellular b1b2 domain of NRP1, which is known to mediate binding to CendR peptides (*13*). The mAb3 was observed to bind to the recombinant b1b2 domain of wild-type (WT) NRP1 but not to the triple-mutant b1b2 domain (S346A, E348A, and T349A in the CendR binding pocket) (fig. S2A). The potency of the mAbs in preventing cellular binding and internalization of NRP ligands was tested using 80-nm silver nanoparticles (AgNP) coated with the prototypic NRP1-binding CendR peptide $RPARPAR_{OH}$ (*9*) (fig. S2B). mAb3 efficiently blocked AgNP-CendR binding (fig. S2C) and internalization (fig. S2, D and E), whereas another monoclonal antibody, mAb2, had no effect and was used as a control in further experiments. Treatment of HEK-293T with mAb3 significantly reduced infection by SARS-CoV-2 pseudoviruses in cells expressing ACE2, TMPRSS2, and NRP1 (Fig. 1D), but not in cells expressing ACE2 and TMPRSS2 only (fig. S2F). When SARS-CoV-2 pseudovirus was preincubated with recombinant, soluble extracellular b1b2 domain of NRP1, the wild type significantly reduced infection compared with the triple mutant (Fig. 1E and fig. S2G).

Next, we explored the role of NRP1 using SARS-CoV-2 isolated from COVID-19 patients from the Helsinki University Hospital. We used WT SARS-CoV-2 and a cleavage-impaired SARS-CoV-2 mutant that was isolated from Vero-E6 cells, which rapidly accumulate mutations at the furin cleavage site of the S protein during passaging (Fig. 2, A and B) (*14*). First, we confirmed that furin cleaved the WT, but not the mutant, SARS-CoV-2 S protein by analyzing S protein processing in Chinese hamster ovary cells with functional (parental) or deficient (FD11) furin enzyme (fig. S3) (*15*). Next, we

[1]Institute of Neuronal Cell Biology, Technical University Munich, Munich, Germany. [2]German Center for Neurodegenerative Diseases (DZNE), Munich, Germany. [3]Faculty of Biological and Environmental Sciences, Molecular and Integrative Biosciences Research Program, University of Helsinki, Helsinki, Finland. [4]Department of Neuropathology, University Medical Center Göttingen, Göttingen, Germany. [5]Campus Institute for Dynamics of Biological Networks, University of Göttingen, Göttingen, Germany. [6]Max Planck Institute for Experimental Medicine, Göttingen, Germany. [7]Department of Virology, Medicum, University of Helsinki, Helsinki, Finland. [8]Institute for Stroke and Dementia Research (ISD), University Hospital, LMU Munich, Munich, Germany. [9]Helsinki Institute of Life Sciences–Institute of Biotechnology, University of Helsinki, Helsinki, Finland. [10]Laboratory of Cancer Biology, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia. [11]Department of Virology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland. [12]Department of Health Security, Finnish Institute for Health and Welfare (THL), Helsinki, Finland. [13]Clem Jones Centre for Ageing Dementia Research, Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia. [14]Department of Anesthesiology and Intensive Care Medicine, University Medical Center Göttingen, Göttingen, Germany. [15]Department of Neurology, University Medical Center Göttingen, Göttingen, Germany. [16]Paracelsus-Elena-Klinik Kassel, Kassel, Germany. [17]Institute of Biochemistry, ETH Zürich, Zürich, Switzerland. [18]The Queensland Brain Institute, The University of Queensland, Brisbane, Queensland, Australia. [19]Cancer Research Center, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, USA. [20]Center for Nanomedicine and Department of Molecular, Cellular, and Developmental Biology, University of California, Santa Barbara, Santa Barbara, CA, USA. [21]Institute of Veterinary Pathology, Vetsuisse Faculty, University of Zürich, Zürich, Switzerland. [22]Department of Veterinary Biosciences, University of Helsinki, Helsinki, Finland. [23]Munich Cluster of Systems Neurology (SyNergy), Munich, Germany.

*These authors contributed equally to this work.

†Corresponding author. Email: giuseppe.balistreri@helsinki.fi (G.B.); mikael.simons@dzne.de (M.S.)

Corrected 12 November 2020. See full text.

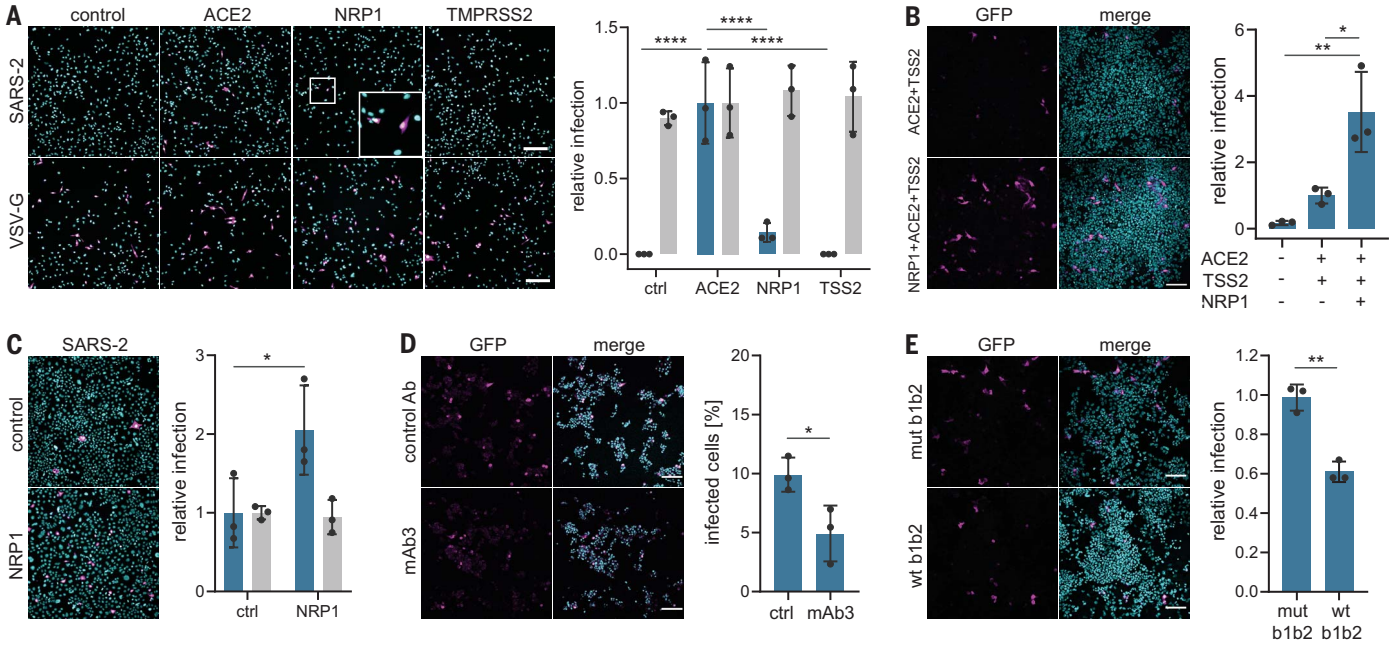


Fig. 1. NRP1 facilitates the cellular entry of SARS-CoV-2 pseudotyped particles. (**A**) Representative images and quantification of SARS-CoV-2 S protein (SARS-2) (blue bars) and vesicular stomatitis virus glycoprotein (VSV-G) pseudotype (gray bars) infectivity in HEK-293T cells transiently expressing control (ctrl) vector, ACE2, NRP1, or TMPRSS2 (TSS2). Data are normalized to the respective infectivity of SARS-2 and VSV-G pseudotype in ACE2-expressing cells. Two-way analysis of variance (ANOVA) was carried out with Tukey's correction for multiple comparisons. (**B**) HEK-293T cells transiently expressing ACE2 and TMPRSS2 or NRP1, ACE2, and TMPRSS2 were inoculated with SARS-2 pseudotype. Data are normalized to SARS-2 infectivity in cells expressing ACE2 and TMPRSS2. One-way ANOVA was performed with Tukey's correction for multiple comparisons. (**C**) SARS-2 pseudotype infectivity in Caco-2 cells expressing NRP1 or control vector. Data are normalized to the respective infectivity of SARS-2 and VSV-G pseudotype in control cells. Two-way ANOVA was carried out with Sidak's correction for multiple comparisons. (**D** and **E**) HEK-293T cells transiently expressing NRP1, ACE2, and TMPRSS2 were inoculated with SARS-2 pseudotype in the presence of mAb3 antibody against NRP1 [(D), mAb3] or control mAb2 [(D), ctrl Ab] and in the presence of soluble NRP1 wild-type b1b2 domain [(E), wt b1b2] or NRP1 mutant b1b2 domain [(E), mut b1b2]. Data in (E) are normalized to untreated cells expressing NRP1, ACE2, and TMPRSS2. Two-tailed unpaired Student's *t* test was performed. All data are represented as means ± SDs from three independent experiments [(A) to (C)] or three biological replicates [(D) and (E)]. $*P < 0.05$, $**P < 0.01$, $****P < 0.0001$. All images show GFP-positive, infected cells (magenta) and Hoechst stain (cyan). Scale bars, 100 μm.

validated that exogenous ACE2 expression rendered HEK-293T cells susceptible to infection with SARS-CoV-2 (Fig. 2, C and D). NRP1 expression alone caused lower levels of infection, which were detectable only with increasing virus titer (Fig. 2, C and D). We then compared the ability of WT and mutant SARS-CoV-2 to infect HEK-293T that stably express ACE2; ACE2 and TMPRSS2; or ACE2, TMPRSS2, and NRP1. Infection of these cell lines by the WT, but not the mutant, virus increased in the presence of NRP1, providing further evidence that NRP1 requires a furin-cleaved substrate for its effects (Fig. 2, E and F). We studied the effect of the NRP1-blocking antibody, mAb3, on infection of Caco-2 cells by WT and mutant SARS-CoV-2 and found that preincubation with NRP1-blocking antibody reduced WT virus infection by ~40%, whereas the control mAb2 had no effect (Fig. 2, G and H). NRP1-blocking antibody had no effect on the infection by the mutated virus (Fig. 2, G and H).

Cleavage of SARS-CoV-2 S protein at the S1-S2 site generates the C-terminal end sequence $TQTNSPRRAR_{OH}$. To determine whether this specific sequence can function as a substrate for NRP1, we used AgNPs coated with $TQTNSPRRAR_{OH}$ peptide or different control peptides, including one with a terminal amide group ($TQTNSPRRAR_{NH2}$), which reduces NRP1 binding (*9*) (Fig. 3A). We found that AgNP-$TQTNSPRRAR_{OH}$, but not control AgNPs, were efficiently taken up by HEK-293T cells expressing NRP1 (Fig. 3, B and C). Next, we determined whether AgNP-$TQTNSPRRAR_{OH}$ particles were also internalized into cells in vivo. We chose to study nanoparticle entry in the mouse olfactory epithelium, owing to the known expression of NRP1 in the olfactory system (*16*), including olfactory neuronal cells of the epithelium (fig. S4). AgNPs-$TQTNSPRRAR_{OH}$ and control AgNP-$TQTNSPRRAR_{NH2}$ were administered into the nose of anesthetized adult mice. Six hours after administration, we observed a significantly larger uptake of AgNP-$TQTNSPRRAR_{OH}$ than of AgNP-$TQTNSPRRAR_{NH2}$ into the olfactory epithelium (Fig. 3, D and E) and, unexpectedly, into neurons and blood vessels of the cortex (Fig. 3, F and G). Similar results were obtained for AgNPs coated with the prototypic NRP1-binding CendR peptide RPAR-PAR_{OH} (fig. S5).

Having obtained evidence for a role of NRP1 in cell entry of SARS-CoV-2, we examined whether NRP1 expression correlated with the detection of virus RNA in single-cell transcriptomes. For these analyses, we used published single-cell RNA sequencing (scRNA-seq) datasets of cultured experimentally infected human bronchial epithelial cells and cells isolated from bronchoalveolar lavage fluid (BALF) of severely affected COVID-19 patients (*17*). Among the proposed entry and amplification factors, *NRP1*, *FURIN*, and *TMPRSS11A* were enriched in SARS-CoV-2–infected cells compared with noninfected cells (fig. S6). We also detected increased expression of these proteins after infection (fig. S6). In addition, RNA expression of *NRP1* and its homolog *NRP2* was elevated in SARS-CoV-2–positive cells compared with adjacent cells in the BALF of severely affected COVID-19 patients (fig. S7).

Because the availability of virus receptors and entry cofactors on the surface of host cells determines infectivity, we compared the expression patterns of *ACE2* and *NRP1* in published scRNA-seq datasets of human lung tissue (*18*) and human olfactory epithelium (*19*). Whereas *ACE2* was detected at very low levels, both *NRP1* and *NRP2* were abundantly expressed in almost all pulmonary and olfactory

Corrected 12 November 2020. See full text.

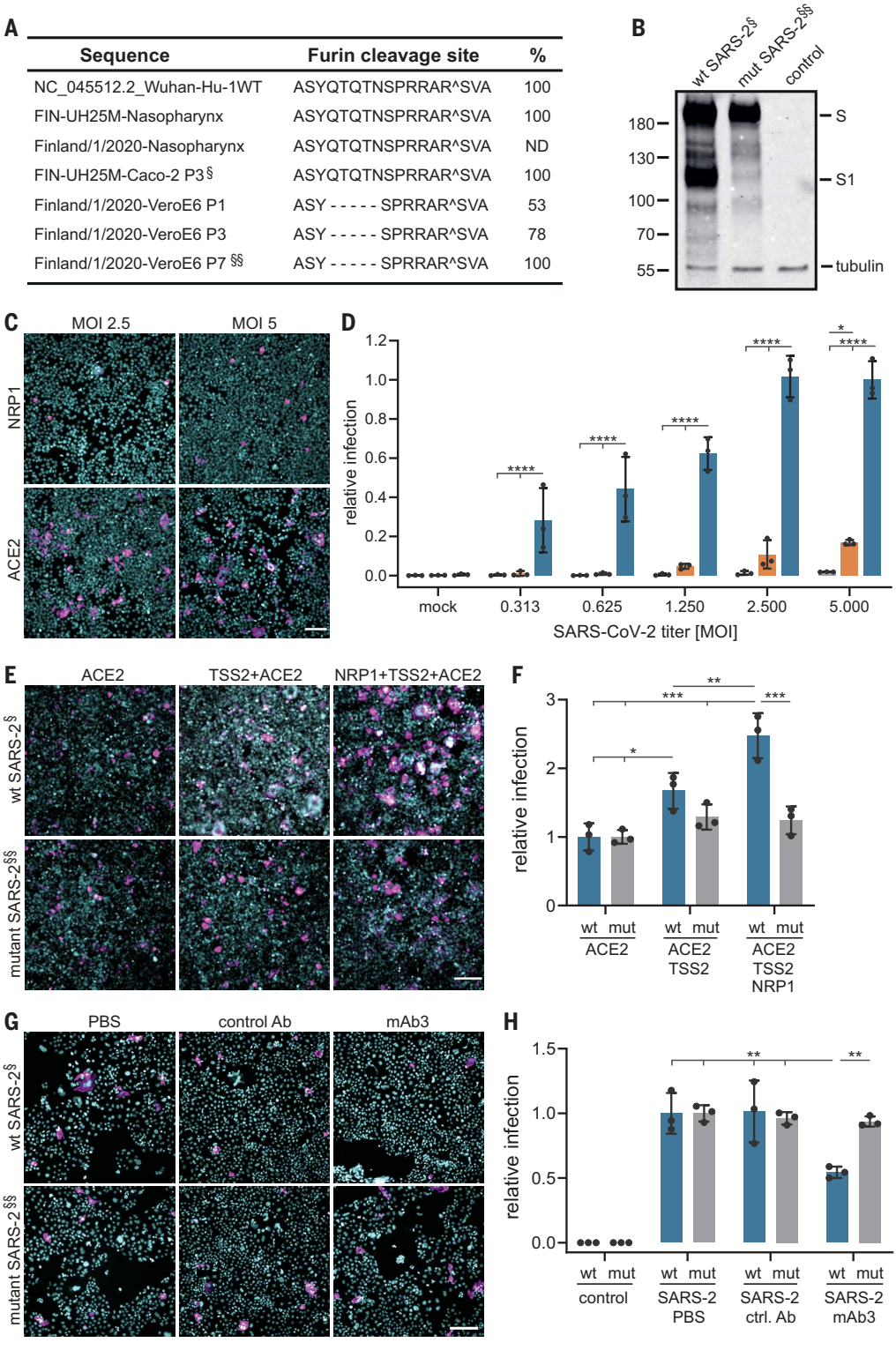


Fig. 2. A blocking antibody against the b1b2 domain of NRP1 reduces infection by wild-type SARS-CoV-2 (SARS-2-wt) but not a mutant with a deletion at the furin-cleavage site (SARS-2-mut). (**A**) Sequence analysis of viruses isolated at different passages (P) from different cell types. The first sequence is the reference from the Wuhan isolate (NC_045512.2). The sequence abundance in each virus population is indicated as a percentage. ND, not determined. §, SARS-2-wt; §§, SARS-2-mut. A, Ala; S, Ser; Y, Tyr; Q, Gln; T, Thr; N, Asn; P, Pro; R, Arg; V, Val. (**B**) A deletion adjacent to the furin-cleavage site abrogates the enzymatic cleavage of the S protein. Immunoblot analysis was carried out on cell lysates from Vero-E6 cells infected for 16 hours with two viral populations (§ and §§). Numbers indicate protein size (in kilodaltons). (**C** and **D**) Representative images and quantification of SARS-2-wt infectivity in HEK-293T cells stably expressing ACE2 (blue bars) or NRP1 (orange bars) compared with nontransfected cells (gray bars). Different virus titers were used. Data are normalized to the infectivity in ACE2-expressing cells at multiplicity of infection (MOI) = 5. Two-way ANOVA was done with Tukey's correction for multiple comparisons. (**E** and **F**) Representative images (E) and quantification (F) of HEK-293T cells stably expressing the indicated combinations of ACE2, TMPRSS2 (TSS2), and NRP1 after inoculation with SARS-2-wt (wt; blue bars) or SARS-2-mut (mut; gray bars). Data are normalized to the respective infectivity in ACE2-expressing cells. Two-way ANOVA was performed with Tukey's correction for multiple comparisons. (**G** and **H**) Caco-2 cell infection in the presence of control mAb2 (ctrl. Ab) or mAb3 blocking antibodies against NRP1 after SARS-2-wt (wt; blue bars) or SARS-2-mut (mut; gray bars) inoculation. Data are normalized to the respective vehicle control (phosphate-buffered saline) sample. Two-way ANOVA was performed with Tukey's correction for multiple comparisons. Data are means ± SDs from three independent experiments. *P < 0.05, **P < 0.01, ***P < 0.001, ****P < 0.0001. Magenta, SARS-2-wt– and SARS-2-mut–infected cells; Hoechst stain, cyan. Scale bars, 50 μm.

cells, with the highest levels of expression in endothelial cells (figs. S8 and S9). We confirmed these results by examining NRP1 immunoreactivity in human autopsy tissue and detected NRP1 in the epithelial surface layer of the human respiratory and olfactory epithelium (fig. S10A). ACE2 was hardly detectable in these tissues (fig. S10B). Within the olfactory epithelium, NRP1 was also observed in cells positive for oligodendrocyte transcription factor 2 (OLIG2), which is mostly expressed by olfactory neuronal progenitors (fig. S10C).

In light of the widely reported disturbance of olfaction in a large fraction of COVID-19

Corrected 12 November 2020. See full text.

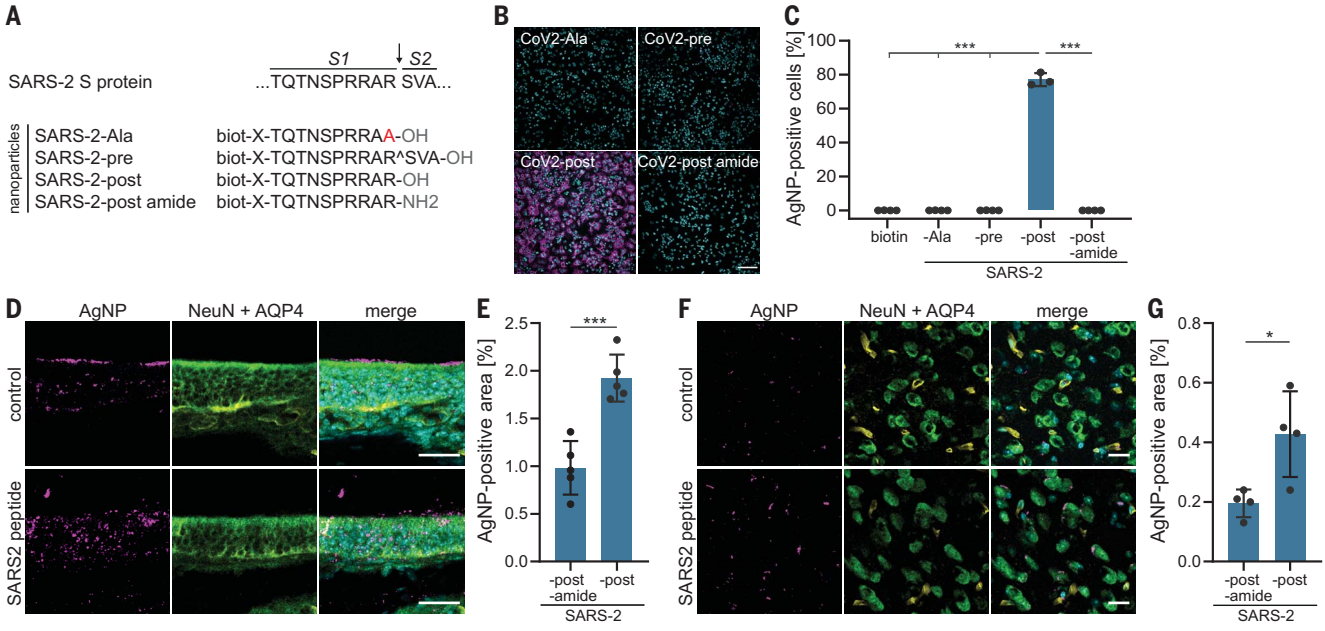


Fig. 3. NRP mediates entry of nanoparticles coated with SARS-CoV-2 (SARS-2) S–derived CendR peptides into cultured cells, olfactory epithelium, and the central nervous system of mice. (**A**) Peptide sequences used for AgNP coating. The peptides mimic SARS-2 S protein after furin cleavage (post) and as controls; S protein before cleavage (pre), in which the terminal amino acid is replaced by alanine (Ala); or with an amide terminus (post amide). X, any amino acid. The arrow indicates the cleavage site. (**B** and **C**) Representative images and quantification of the internalization of peptide-coated AgNPs in HEK-293T cells expressing NRP1. Merged images show AgNP-positive cells (magenta) and Hoechst stain (cyan). One-way ANOVA was carried out with Tukey's correction for multiple comparisons. (**D** to **G**) Representative images and quantification of the main olfactory epithelium [(D) and (E), respectively] and cortex [(F) and (G), respectively] 6 hours after intranasal administration of AgNPs coated with SARS2-post and SARS2-post amide peptides. n = 4 replicates for (C); n = 5 (E) and n = 4 (G) mice per condition. Data are means ± SDs. Two-tailed unpaired Student's t test; $^{*}P < 0.05$, $^{***}P < 0.001$. Magenta, AgNPs; cyan, Hoechst stain; green, NeuN (neuronal marker); yellow, AQP4. Scale bars, 100 μm (B), 20 μm [(D) and (F)].

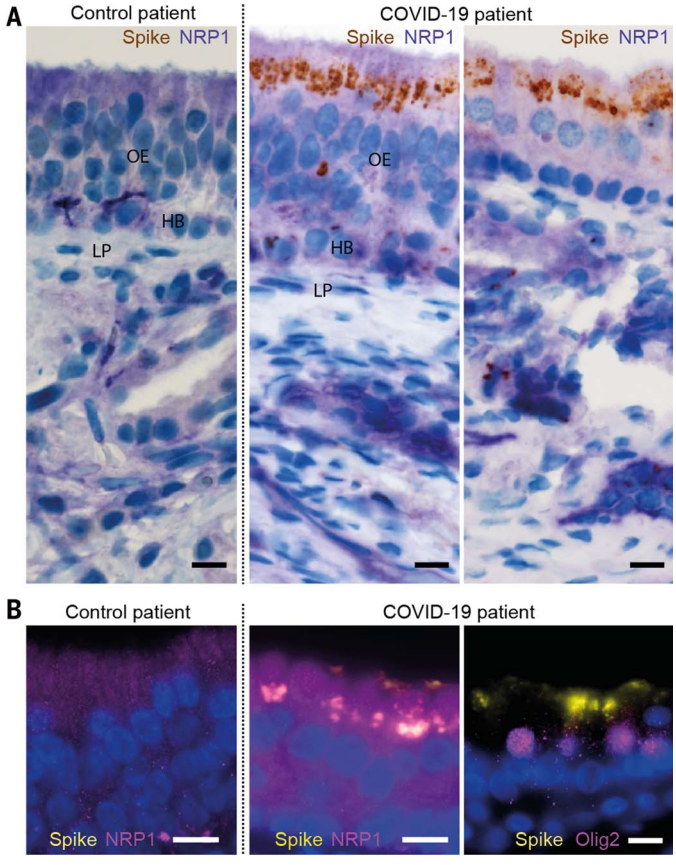


Fig. 4. SARS-CoV-2 infects the olfactory epithelium. (**A**) Costaining of S protein (brown) and NRP1 (purple) in the apical olfactory epithelium (OE) in a noninfected control (left) and in the apical OE (middle) and adjacent mucosa (right) in a COVID-19 patient. LP, lamina propria; HB, horizontal basal cells. (**B**) Costaining of NRP1 (magenta) or OLIG-2 (magenta) with S protein (yellow) in OE cells. Nuclei are shown in blue. Scale bars, 10 μm.

patients (*20*, *21*) and the enrichment of NRPs in the olfactory epithelium, we analyzed a series of autopsies from six COVID-19 patients and eight noninfected control patients to determine whether SARS-CoV-2 could infect NRP1-positive cells (Fig. 4 and table S1). Using antibodies against the S protein, we detected infection of the olfactory epithelium in five of six COVID-19 patients. The infected olfactory epithelial cells showed high expression of NRP1 (Fig. 4, A and B). Additional costaining indicated infection of cells positive for OLIG2 (Fig. 4B and fig. S11).

There is limited knowledge about the virus–host interactions that determine cellular entry of SARS-CoV-2. Viruses display considerable redundancy and flexibility because they can exploit weak multivalent interactions to enhance affinity. To date, studies of SARS-CoV-2 entry have focused almost entirely on ACE2, which is expressed at very low protein levels in respiratory and olfactory epithelial cells (*22*). This raises the possibility that cofactors are required to facilitate virus–host cell interactions in cells with low ACE2 expression. NRP1 could represent such an ACE2 potentiating factor by promoting the interaction of the virus with ACE2. The reason a number of viruses (*23*–*26*) use NRPs as entry factors may be their high expression on epithelia facing the external environment and their function in

Corrected 12 November 2020. See full text.

enabling cell, vascular, and tissue penetration (*9*, *13*).

REFERENCES AND NOTES

1. N. Zhu *et al*., *N. Engl. J. Med.* **382**, 727–733 (2020).
2. R. Wölfel *et al*., *Nature* **581**, 465–469 (2020).
3. P. Zhou *et al*., *Nature* **579**, 270–273 (2020).
4. M. Hoffmann *et al*., *Cell* **181**, 271–280.e8 (2020).
5. N. J. Matheson, P. J. Lehner, *Science* **369**, 510–511 (2020).
6. B. Coutard *et al*., *Antiviral Res.* **176**, 104742 (2020).
7. L. V. Tse, A. M. Hamilton, T. Friling, G. R. Whittaker, *J. Virol.* **88**, 1673–1683 (2014).
8. M. Hoffmann, H. Kleine-Weber, S. Pöhlmann, *Mol. Cell* **78**, 779–784.e5 (2020).
9. T. Teesalu, K. N. Sugahara, V. R. Kotamraju, E. Ruoslahti, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16157–16162 (2009).
10. D. Wrapp *et al*., *Science* **367**, 1260–1263 (2020).
11. A. C. Walls *et al*., *Cell* **181**, 281–292.e6 (2020).
12. K. Liao, D. Sikkema, C. Wang, T. N. Lee, *J. Immunol. Methods* **389**, 52–60 (2013).
13. A. Plein, A. Fantin, C. Ruhrberg, *Microcirculation* **21**, 315–323 (2014).
14. A. D. Davidson *et al*., *Genome Med.* **12**, 68 (2020).
15. N. G. Ravindra *et al*., bioRxiv 2020.05.06.081695 [Preprint]. 13 July 2020.
16. A. Kawakami, T. Kitsukawa, S. Takagi, H. Fujisawa, *J. Neurobiol.* **29**, 1–17 (1996).
17. M. Liao *et al*., *Nat. Med.* **26**, 842–844 (2020).
18. X. Han *et al*., *Nature* **581**, 303–309 (2020).
19. M. A. Durante *et al*., *Nat. Neurosci.* **23**, 323–326 (2020).
20. L. Mao *et al*., *JAMA Neurol.* **77**, 683–690 (2020).
21. K. W. Cooper *et al*., *Neuron* **107**, 219–233 (2020).
22. F. Hikmet *et al*., *Mol. Syst. Biol.* **16**, e9610 (2020).
23. D. Ghez *et al*., *J. Virol.* **80**, 6844–6854 (2006).
24. N. Martinez-Martin *et al*., *Cell* **174**, 1158–1171.e19 (2018).
25. H. B. Wang *et al*., *Nat. Commun.* **6**, 6240 (2015).
26. M. Raaben *et al*., *Cell Host Microbe* **22**, 688–696.e5 (2017).

ACKNOWLEDGMENTS

We thank R. Müller, K. Schulz, and U. Scheidt for expert technical assistance; S. Osborne for proofreading the manuscript, and the DNA Dream Lab facility and K. Kogan for design and cloning of plasmids. **Funding:** The work in Munich and Göttingen was supported by grants from the German Research Foundation (SPP2191, TRR128-2, TRR274-1, SyNergy Excellence Cluster, EXC2145, Projekt ID390857198, EXC 2067/1- 390729940, and STA 1389/5-1), the ERC (Consolidator Grant to M.S.), and the Dr. Miriam and Sheldon G. Adelson Medical Research Foundation. The work at the University of Helsinki was supported by the University of Helsinki and by donations from Finnish colleagues to whom we are very grateful. The Academy of Finland supported G.B. (318434), O.V. (336490), S.J.B. (315950 and 336471), and J.H. (1308613 and 1314119). O.V. was supported by Jane and Aatos Erkko Foundation, EU Horizon 2020 program VEO (874735), and Helsinki University Hospital Funds (TYH2018322). S.J.B. was supported by the Swedish Research Foundation and M.A. by the Marie Sklodowska-Curie Actions (799929). M.J. is supported by The Australian Research Council's Discovery Early Career Researcher Award (DE190100565). F.A.M. is supported by an Australian National Health and Medical Research Council Senior Research Fellowship (GNT1155794). T.T. and A.T. are supported by the European Regional Development Fund (project 2014-2020.4.01.15-0012), Wellcome Trust International Fellowship WT095077MA, European Research Council grant GLIOGUIDE, and the Estonian Research Council (grants PRG230 and EAG79 to T.T.). **Author contributions:** G.B., M.S., and A.H. conceived the project. L.C.C., R.O., L.D.P., M.D., J.F., S.K., F.v.d.M., K.K., M.A., and L.S. designed and carried out experiments. A.T., T.T., L.L., O.V., J.H., O.G., H.K.K., P.O., and M.J., developed and provided tools. L.C.C., R.O., L.D.P., M.D., J.F., S.K., T.K., C.S., T.S., M.J., F.A.M., S.J.B., J.H., and O.V. analyzed the data or supervised data acquisition. L.C.C., R.O., L.D.P., M.D., J.F., S.K., T.K., and O.G. visualized the data. T.K. and O.G. performed the scRNA-seq data analysis. M.S.W., B.M., C.S., and H.K.K. provided human samples. G.B. and M.S. wrote the manuscript. G.B. and M.S. supervised the project. **Competing interests:** T.T., O.V., and G.B. have a pending patent on the monoclonal antibody 3 (mAb3) against the NRP1 b1 domain for SARS-CoV-2 inhibition. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials.
This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/856/suppl/DC1
Materials and Methods
Figs. S1 to S11
Table S1
References (*27*–*37*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

12 June 2020; accepted 12 October 2020
Published online 20 October 2020
10.1126/science.abd2985

CORONAVIRUS

Neuropilin-1 is a host factor for SARS-CoV-2 infection

James L. Daly[1]*, Boris Simonetti[1]*†, Katja Klein[2]*, Kai-En Chen[3]‡, Maia Kavanagh Williamson[2]‡, Carlos Antón-Plágaro[1]‡, Deborah K. Shoemark[4], Lorena Simón-Gracia[5], Michael Bauer[6], Reka Hollandi[7], Urs F. Greber[6], Peter Horvath[7,8], Richard B. Sessions[1], Ari Helenius[9], Julian A. Hiscox[10,11], Tambet Teesalu[5], David A. Matthews[2], Andrew D. Davidson[2], Brett M. Collins[3], Peter J. Cullen[1]†, Yohei Yamauchi[2,12]†

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of coronavirus disease 2019 (COVID-19), uses the viral spike (S) protein for host cell attachment and entry. The host protease furin cleaves the full-length precursor S glycoprotein into two associated polypeptides: S1 and S2. Cleavage of S generates a polybasic Arg-Arg-Ala-Arg carboxyl-terminal sequence on S1, which conforms to a C-end rule (CendR) motif that binds to cell surface neuropilin-1 (NRP1) and NRP2 receptors. We used x-ray crystallography and biochemical approaches to show that the S1 CendR motif directly bound NRP1. Blocking this interaction by RNA interference or selective inhibitors reduced SARS-CoV-2 entry and infectivity in cell culture. NRP1 thus serves as a host factor for SARS-CoV-2 infection and may potentially provide a therapeutic target for COVID-19.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the coronavirus responsible for the current coronavirus disease 2019 (COVID-19) pandemic (*1*, *2*). A marked difference between the spike (S) protein of SARS-CoV-2 and SARS-CoV is the presence, in the former, of a polybasic sequence motif, Arg-Arg-Ala-Arg (RRAR), at the S1/S2 boundary. It provides a cleavage site for a host proprotein convertase, furin (*3*–*5*) (fig. S1A). The resulting two proteins, S1 and S2, remain noncovalently associated, with the serine protease TMPRSS2 further priming S2 (*6*). Furin-mediated processing increases infectivity and affects the tropism of SARS-CoV-2, whereas furin inhibition diminishes SARS-CoV-2 entry, and deletion of the polybasic site in the S protein reduces syncytia formation in cell culture (*3*–*5*, *7*).

The C terminus of the S1 protein generated by furin cleavage has an amino acid sequence (^{682}RRAR^{685}) that conforms to a [R/K]XX[R/K] motif, termed the "C-end rule" (CendR) (fig. S1B) (*8*). CendR peptides bind to neuropilin-1 (NRP1) and NRP2, transmembrane receptors that regulate pleiotropic biological processes, including axon guidance, angiogenesis, and vascular permeability (*8*–*10*). To explore the possibility that the SARS-CoV-2 S1 protein may associate with neuropilins, we generated a green fluorescent protein (GFP)–tagged S1 construct (GFP-S1) (fig. S1C). When expressed in human embryonic kidney 293T (HEK293T) cells engineered to express the SARS-CoV-2 receptor angiotensin-converting enzyme 2 (ACE2), GFP-S1 immunoprecipitated endogenous NRP1 and ACE2 (Fig. 1A). We transiently coexpressed NRP1-mCherry and either GFP-S1 or GFP-S1 ΔRRAR (a deletion of the terminal ^{682}RRAR^{685} residues) in HEK293T cells. NRP1 immunoprecipitated the S1 protein, and deletion of the CendR motif reduced this association (Fig. 1B). Comparable binding was also observed with mCherry-NRP2, a receptor with high homology to NRP1 (fig. S1, D and E). In both cases, residual binding was observed with the ΔRRAR mutant, indicating an additional CendR-independent association between neuropilins and the S1 protein.

To probe the functional relevance of this interaction, we generated HeLa wild-type and NRP1 knockout (KO) cell lines stably expressing ACE2, designated as HeLa^{wt}+ACE2 and HeLa^{NRP1KO}+ACE2, respectively (the level of ACE2 expression was comparable between these lines) (fig. S1F). Using a clinical isolate SARS-CoV-2 (SARS-CoV-2/human/Liverpool/REMRQ001/2020), we performed viral infection assays and fixed the cells at 6 and 16 hours postinfection (hpi). SARS-CoV-2 infection was reduced in HeLa^{NRP1KO}+ACE2 relative to HeLa^{wt}+ACE2 (Fig. 1C). HeLa cells lacking ACE2 expression were not infected (fig. S1G). In Caco-2 cells, a human colon adenocarcinoma cell line endogenously expressing ACE2 and widely used in COVID-19 studies, the suppression of NRP1 expression by short hairpin RNA (shRNA) greatly reduced SARS-CoV-2 infection at both 7 and 16 hpi, respectively, whereas that of vesicular stomatitis virus (VSV) pseudotyped with VSV-G was unaffected (Fig. 1D and figs. S1H and S2A). To determine if NRP1 was required for early virus infection, we established a sequential staining procedure using antibodies against SARS-CoV-2 S and N proteins to distinguish extracellular and intracellular viral particles (fig. S2B). Although NRP1 depletion did not affect SARS-CoV-2 binding to the Caco-2 cell surface (Fig. 1E), virus uptake was halved in NRP1-depleted cells compared to control cells after 30 min of internalization (Fig. 1F). Thus, NRP1 enhances SARS-CoV-2 entry and infection.

We also observed that SARS-CoV-2–infected HeLa^{wt}+ACE2 cells displayed a multinucleated syncytia cell pattern, as reported by others (Fig. 1C) (*5*). Using an image analysis algorithm and supervised machine learning (fig. S2, C to F) (*11*), we quantified syncytia of infected HeLa^{wt}+ACE2 and HeLa^{NRP1KO}+ACE2 cells. At 16 hpi, the majority of HeLa^{wt}+ACE2 cells formed syncytia, whereas in HeLa^{NRP1KO}+ACE2 cells, this phenotype was reduced (fig. S2G). When infected with a SARS-CoV-2 isolate lacking the furin cleavage site (SARS-CoV-2 ΔS1/S2) (fig. S1A), the differences in infection and syncytia formation were less pronounced (fig. S2, H and I). However, a significant decrease in infection of HeLa^{NRP1KO}+ACE2 was still observed at 16 hpi, indicating that NRP1 may additionally influence infection through a CendR-independent mechanism (fig. S2H).

The extracellular regions of NRP1 and NRP2 are composed of two CUB domains (a1 and a2), two coagulation factor domains (b1 and b2), and a MAM domain (*9*). Of these, the b1 domain contains the specific binding site for CendR peptides (fig. S3A) (*12*). Accordingly, the mCherry-b1 domain of NRP1 immunoprecipitated GFP-S1, and a shortened GFP-S1 construct spanning residues 493 to 685 (figs. S1C and S3B). Isothermal titration calorimetry (ITC) established that the b1 domain of NRP1 directly bound a synthetic S1 CendR peptide (^{679}NSPRRAR^{685}) with an affinity of 20.3 μM at pH 7.5, which was enhanced to 13.0 μM at pH 5.5 (Fig. 2A). Binding was not observed to an S1 CendR peptide in which the C-terminal arginine was mutated to alanine (^{679}NSPRRAA^{685}) (Fig. 2A). We cocrystallized the NRP1 b1 domain in complex with the S1 CendR peptide (Fig. 2B). The resolved 2.35-Å structure revealed four molecules of b1 with electron density of the S1 CendR peptide clearly visible in the asymmetric unit (fig. S3C). S1 CendR peptide binding displayed strong similarity to the previously solved structure of NRP1 b1 domain in complex with

[1]School of Biochemistry, Faculty of Life Sciences, Biomedical Sciences Building, University of Bristol, Bristol BS8 1TD, UK. [2]School of Cellular and Molecular Medicine, Faculty of Life Sciences, Biomedical Sciences Building, University of Bristol, Bristol BS8 1TD, UK. [3]Institute for Molecular Bioscience, the University of Queensland, St. Lucia, QLD 4072, Australia. [4]School of Biochemistry and BrisSynBio Centre, Faculty of Life Sciences, Biomedical Sciences Building, University of Bristol, Bristol BS8 1TD, UK. [5]Laboratory of Cancer Biology, Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia. [6]Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland. [7]Synthetic and Systems Biology Unit, Biological Research Centre (BRC), Szeged, Hungary. [8]Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. [9]Institute of Biochemistry, ETH Zurich, Zurich, Switzerland. [10]Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK. [11]Singapore Immunology Network, Agency for Science, Technology, and Research, 138648, Singapore. [12]Division of Biological Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan.
*These authors contributed equally to this work.
†Corresponding author. Email: bs13866@bristol.ac.uk (B.S.); pete.cullen@bristol.ac.uk (P.J.C.); yohei.yamauchi@bristol.ac.uk (Y.Y.)
‡These authors contributed equally to this work.

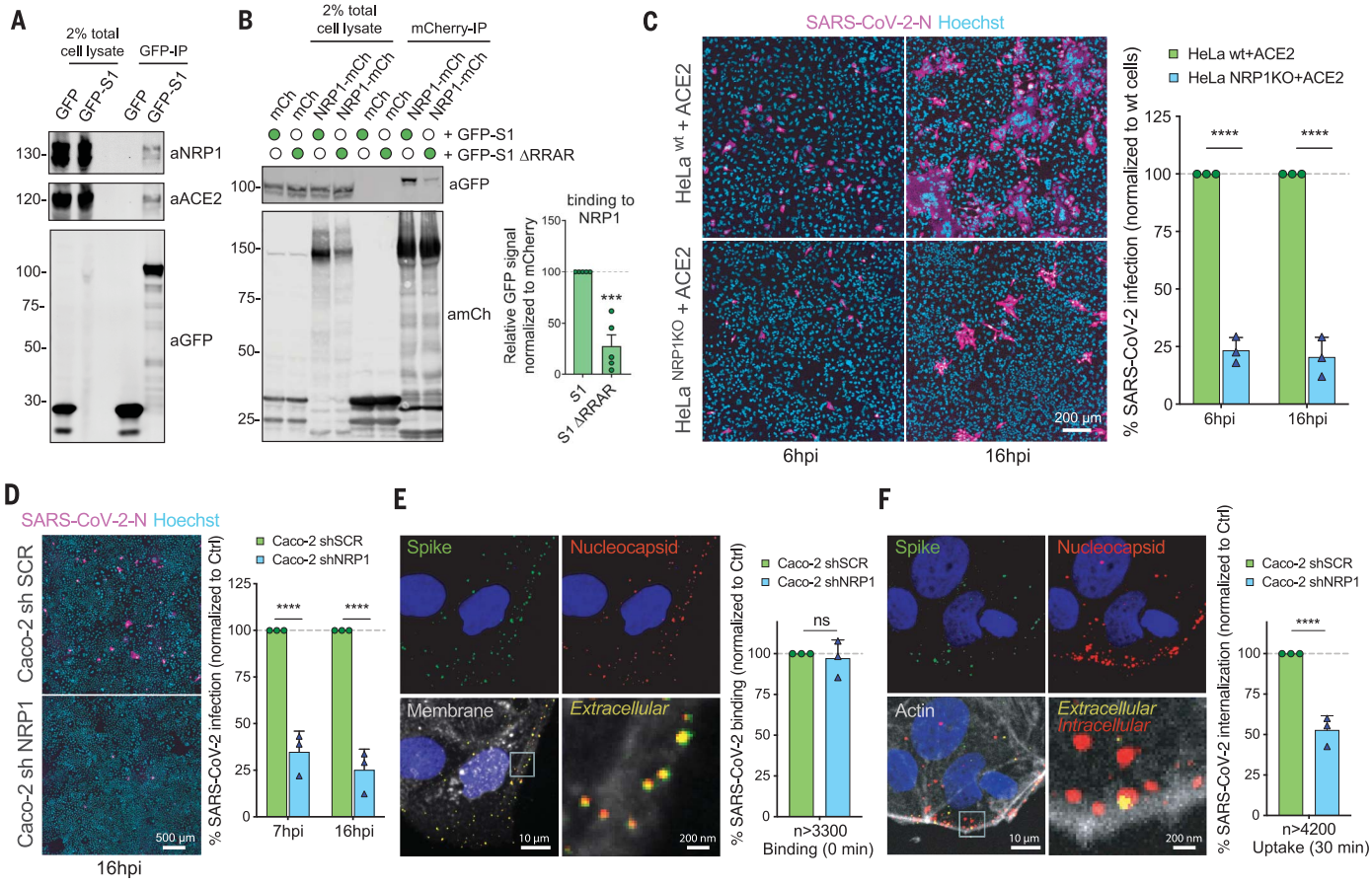


Fig. 1. NRP1 Interacts with S1 and enhances SARS-CoV-2 infection. (**A**) HEK293T cells transduced to express ACE2 were transfected to express GFP or GFP-tagged S1 and lysed after 24 hours. The lysates were subjected to GFP-nanotrap, and the immune isolates were blotted for ACE2 and NRP1 (N = 3 independent experiments). (**B**) HEK293T cells were cotransfected to express GFP-tagged S1 or GFP-S1 ΔRRAR and mCherry or mCherry-tagged NRP1 and subjected to GFP-nanotrap (N = 5 independent experiments). Two-tailed unpaired t test; P = 0.0002. (**C**) HeLa^{wt}+ACE2 and HeLa$^{NRP1\ KO}$+ACE2 cells were infected with SARS-CoV-2. Cells were fixed at 6 or 16 hpi and stained for N protein (magenta) and Hoechst (cyan), and virus infectivity was quantified (N = 3 independent experiments). Two-tailed unpaired t test; P = 0.00002 and 0.00088. Scale bar, 200 µm. (**D**) Caco-2 cells expressing shRNA against NRP1 or a nontargeting control (SCR) were infected with SARS-CoV-2 and fixed at 7 or 16 hpi. The cells were stained for N protein (magenta) and Hoechst (cyan), and infectivity was quantified (N = 3 independent experiments). Two-tailed unpaired t test; P = 0.0005 and 0.00032. Scale bar, 500 µm. (**E**) Caco-2 shSCR or shNRP1 cells were inoculated with a multiplicity of infection (MOI) = 50 of SARS-CoV-2 and incubated in the cold for 60 min, and fixed. A two-step antibody staining procedure was performed with antibodies against S and N to distinguish external (green) and total (red) virus particles, and the binding of particles per cell was quantified for >3300 particles per condition (N = 3 independent experiments). Two-tailed unpaired t test; P = 0.6859. (**F**) Caco-2 shSCR or shNRP1 cells were bound with SARS-CoV-2 as in (E), followed by incubation at 37°C for 30 min. The cells were fixed and stained as in (E). Viral uptake was quantified for >4200 particles per condition (N = 3 independent experiments). Two-tailed unpaired t test; P = 0.00079. Scale bars [(E) and (F)], 10 µm and 200 nm (magnified panels). The square regions were enlarged. The bars, error bars, and circles and triangles represent the mean, SEM (B) and SD [(C) to (F)], and individual data points, respectively. ***P < 0.001, ****P < 0.0001. ns, not signficant.

its endogenous ligand VEGF-A$_{164}$ (Fig. 2B and fig. S3D) (*12*). The key residues responsible for contacting the C-terminal R685 of the CendR peptide —Y297, W301, T316, D320, S346, T349 and Y353—are almost identical between the two structures (Fig. 2B and fig. S3D). The R682 and R685 side chains together engage NRP1 via stacked cation-π interactions with NRP1 side chains of Y297 and Y353. By projecting these findings onto the structure of the NRP1 ectodomain, the b1 CendR binding pocket appears to be freely accessible to the S1 CendR peptide (fig. S3E) (*13*).

Site-directed mutagenesis of the S1 R685 residue to aspartic acid drastically reduced GFP-S1$^{493\text{-}685}$ immunoprecipitation by mCherry-b1, confirming the critical role of the C-terminal arginine (Fig. 2C). Mutagenesis of the T316 residue within the mCherry-b1 domain of NRP1 to arginine also reduced association with GFP-S1$^{493\text{-}685}$, consistent with its inhibitory impact on VEGF-A$_{164}$ binding (*12*) (Fig. 2D). Accordingly, incubation of mCherry-b1 with VSV particles pseudotyped with trimeric S resulted in immunoprecipitation of processed forms of S1, which was dependent on the T316 residue (fig. S3F). Next, we transiently expressed either GFP, full-length NRP1 wt-GFP, or full length NRP1-GFP harboring the T316R mutation in HeLa^{NRP1KO}+ACE2 cells. GFP expression and ACE2 expression levels were comparable and both constructs retained similar cell surface localization (fig. S3, G and H). SARS-CoV-2 infection was significantly enhanced in cells expressing NRP1 wt-GFP compared to GFP control, whereas it was not enhanced in cells expressing the T316R mutant (Fig. 2E). Thus, the SARS-CoV-2 S1 CendR and NRP1 interaction promotes infection.

To establish the functional relevance of the S1 CendR-NRP1 interaction, we screened monoclonal antibodies (mAb#1, mAb#2, mAb#3) raised against the NRP1 b1b2 ectodomain. All three bound to the NRP1 b1b2 domain, displayed staining by immunofluorescence

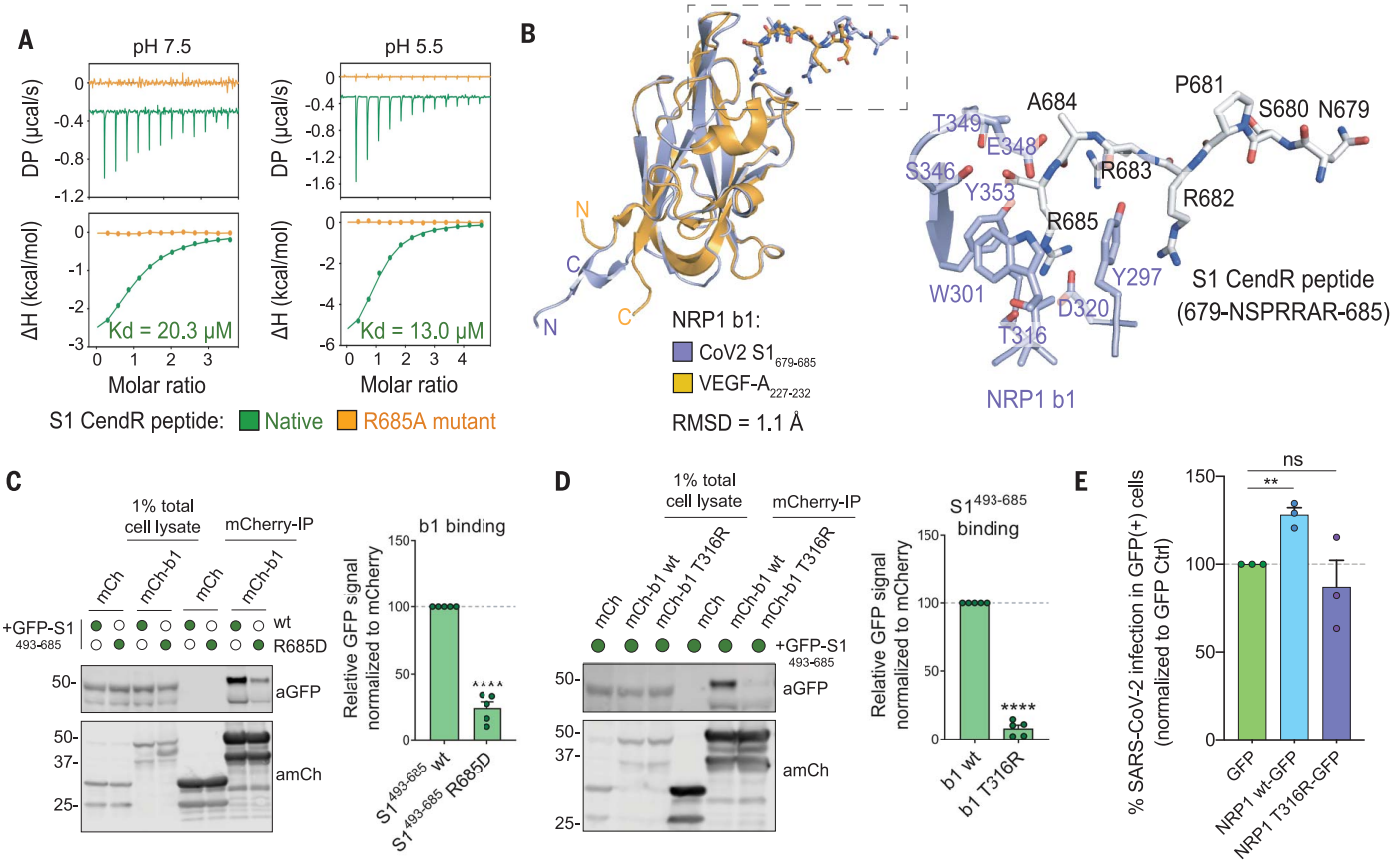


Fig. 2. Molecular basis for CendR binding of SARS-CoV-2 S1 with NRP1. (**A**) Binding of NRP1 b1 with native (green line) and mutant (orange line) form of S1 CendR peptide (corresponding to residues 679 to 685) by ITC at two different pH conditions (N = 3 independent experiments). All ITC graphs represents the integrated and normalized data fit with 1-to-1 ratio binding. (**B**) (Left) NRP1 b1–S1 CendR peptide complex superposed with NRP1 b1–VEGF-A fusion complex (PDB ID: 4DEQ). Bound peptides are shown in stick representation. RMSD, root mean square deviation. (Right) Enlarged view highlighting the binding of S1 CendR peptide b1. Key binding residues on b1 are shown in stick representation. Abbreviations for the amino acid residues are as follows: A, Ala; D, Asp; E, Glu; N, Asn; P, Pro; R, Arg; S, Ser; T, Thr; W, Trp; and Y, Tyr. (**C**). HEK293T cells were cotransfected with combinations of GFP-tagged $S1^{493\text{-}685}$ and $S1^{493\text{-}685}$ R685D, and mCherry or mCherry-NRP1 b1, and subjected to mCherry-nanotrap (N = 5 independent experiments). Two-tailed unpaired t test; $P < 0.0001$. (**D**). HEK293T cells were cotransfected with combinations of GFP-tagged $S1^{493\text{-}685}$ and mCherry, mCherry-NRP1 b1 or mCherry-NRP1 b1 T316R mutant, and subjected to mCherry-nanotrap (N = 5 independent experiments). Two-tailed unpaired t test; $P < 0.0001$. (**E**) $HeLa^{NRP1KO}$ + ACE2 cells transfected with GFP, NRP1 wt-GFP, or NRP1 T316R-GFP constructs were infected 24 hours later with SARS-CoV-2. At 16 hpi, the cells were fixed and stained for SARS-CoV-2-N, and viral infection was quantified in the GFP-positive subpopulation of cells (N = 3 independent experiments). The percentage of infection was normalized to that of GFP-transfected cells. Two-tailed unpaired t test; $P = 0.002$. The bars, error bars, and circles represent the mean, SEM [(C) and (D)] and SD (E), and individual data points, respectively. $^{**}P < 0.01$, $^{****}P < 0.0001$. ns, not signficant.

in NRP1-expressing PPC-1 (human primary prostate cancer) cells but not in M21 (human melanoma) cells that do not express NRP1 (fig. S4A) (*8*), and stained the extracellular domain of NRP1-GFP expressed in cells (fig. S4B). Of these antibodies, mAb#3, and to a lesser extent mAb#1, bound to the CendR-binding pocket with high specificity, as defined by reduced ability to bind to a b1b2 mutant that targets residues (S346, E348, T349) at the opening of the binding pocket (Fig. 3A) (*12*). Incubation of Caco-2 cells with mAbs#1 and 3 reduced SARS-CoV-2 infection compared to a control mAb targeting avian influenza A virus (H11N3) hemagglutinin (Fig. 3B). Consistent with this, mAb#3 inhibited binding of GFP-$S1^{493\text{-}685}$ and mCherry-b1 (Fig. 3C). As a comparison, Caco-2 and Calu-3 cells were incubated with soluble ACE2, which inhibited SARS-CoV-2 infection in both cases (fig. S4C).

Next, we turned to the small molecule EG00229, a selective NRP1 antagonist that binds the b1 CendR binding pocket and inhibits VEGF-A binding (Fig. 3D) (*14*). ITC established that EG00229 bound to the NRP1 b1 domain with a dissociation constant (K_d) of 5.1 and 11.0 µM at pH 7.5 and 5.5, respectively (Fig. 3E). EG00229 inhibited the direct binding between b1 and the S1 CendR peptide, and the immunoprecipitation of GFP-$S1^{493\text{-}685}$ by mCherry-b1 (Fig. 3E and fig. S4D). Finally, incubation of Caco-2 cells with EG00229 reduced the efficiency of SARS-CoV-2 infection at 7 and 16 hpi (Fig. 3F). Thus, the SARS-CoV-2 interaction with NRP1 can be targeted to reduce viral infectivity in relevant human cell lines (fig. S5).

Cell entry of SARS-CoV-2 depends on priming by host cell proteases (*5*, *6*, *15*). Our data indicate that a component of SARS-CoV-2 S protein binding to cell surface neuropilins occurs via the S1 CendR motif generated by the furin cleavage of S1/S2. Though not affecting cell surface attachment, this interaction promotes entry and infection by SARS-CoV-2 in physiologically relevant cell lines widely used in the study of COVID-19. The molecular basis for the effect is unclear, but neuropilins are known to mediate the internalization of CendR ligands through an endocytic process resembling macropinocytosis,

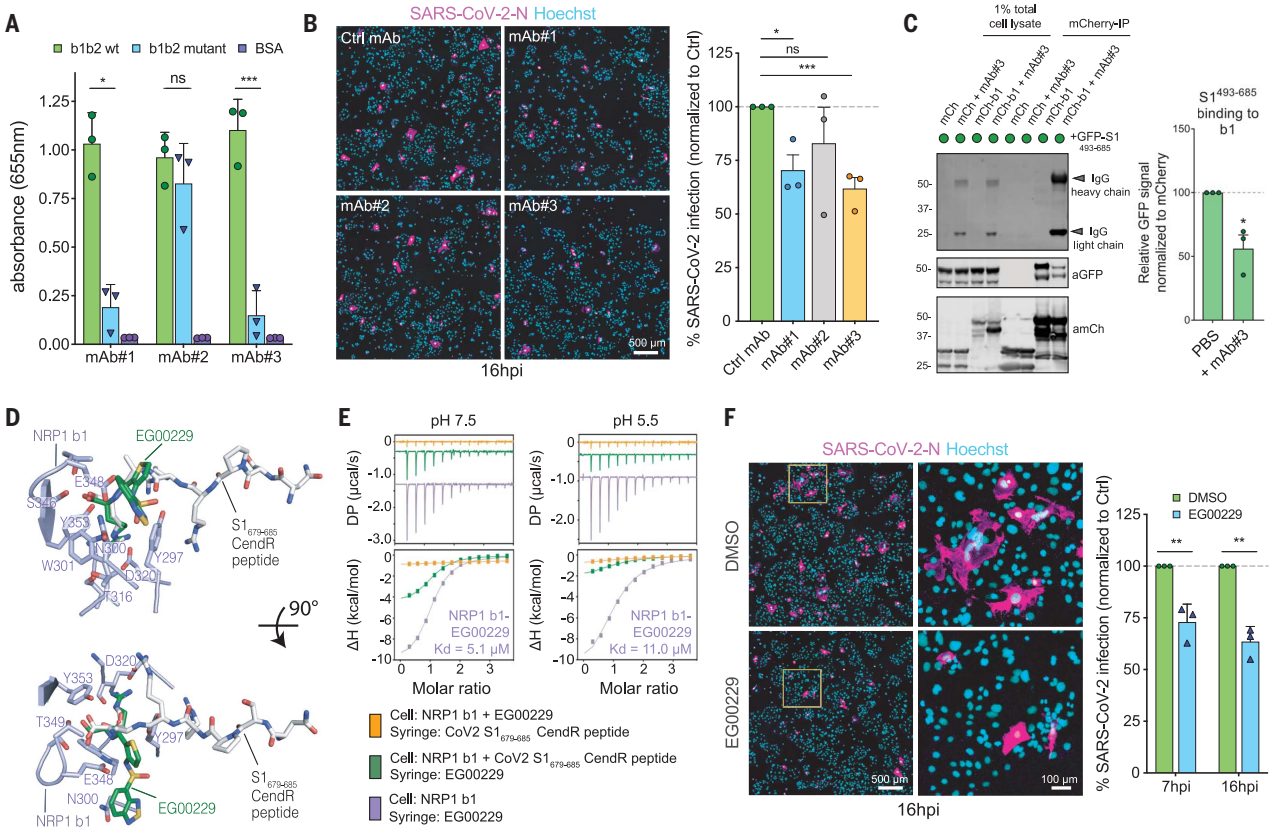


Fig. 3. Selective inhibition of the S1-NRP1 interaction reduces SARS-CoV-2 infection. (**A**) Enzyme-linked immunosorbent assay of anti-NRP1 monoclonal antibodies (mAb#1, mAb#2, mAb#3) at 3 µg/ml using plates coated with NRP1 b1b2 wild type, b1b2 mutant (S346A, E348A, T349A), or bovine serum albumin (BSA), used as a control (N = 3 independent experiments). Binding is represented as arbitrary units of absorbance at 655 nm. Two-tailed unpaired t test; P = 0.0207, 0.2430, 0.0007. (**B**) Cells were first treated with anti-H11N3 (100 µg/ml) (Ctrl) mAb, mAb#1, mAb#2, or mAb#3 for 1 hour before infection with SARS-CoV-2. Cells were fixed at 16 hpi and stained for N protein (magenta) and Hoechst (cyan) (N = 3 independent experiments). Two-tailed unpaired t test; P = 0.015, 0.36, 0.0003. Scale bar, 500 µm. (**C**) HEK293T cells were cotransfected with combinations of mCherry or mCherry-b1 and GFP-tagged $S1^{493-685}$ and subjected to mCherry-nanotrap with or without coincubation with mAb#3 (N = 3 independent experiments). Two-tailed unpaired t test; P = 0.0143. (**D**) NRP1 b1–S1 CendR peptide complex superimposed with NRP1 b1–EG00229 inhibitor complex (PDB ID:3I97). Key binding residues on b1, bound peptides, and EG00229 are shown in stick representation. (**E**) ITC analysis of EG00229 binding to b1 domain of NRP1 at two different pH conditions. Preincubation with EG00229 blocks S1 CendR peptide binding (orange line), and the CendR peptide can reduce binding of EG00229 (green line) (N = 3 independent experiments). All ITC graphs represent the integrated and normalized data fit with 1-to-1 ratio binding. (**F**). Cells were first treated with 100 µM EG00229 or dimethyl sulfoxide before infection with SARS-CoV-2. Cells were fixed at 7 and 16 hpi and stained for N protein (magenta) and Hoechst (cyan) (N = 3 independent experiments). The square regions were enlarged. Scale bars, 500 µm and 100 µm (magnified panels). Two-tailed unpaired t test; P = 0.0059 and 0.0013. The bars, error bars, and circles and triangles represent the mean, SEM (C) and SD [(A), (B), and (F)], and individual data points, respectively. *P < 0.05, **P < 0.01, ***P < 0.001.

(*8*, *16*, *17*). Notably, gene expression analysis has revealed an up-regulation of NRP1 and NRP2 in lung tissue from COVID-19 patients (*18*). A SARS-CoV-2 virus with a natural deletion of the S1/S2 furin cleavage site demonstrated attenuated pathogenicity in hamster models (*19*). NRP1 binding to the CendR peptide in S1 is thus likely to play a role in the increased infectivity of SARS-CoV-2 compared with SARS-CoV. The ability to target this specific interaction may provide a route for COVID-19 therapies.

REFERENCES AND NOTES

1. WHO Coronavirus disease, 2019 (COVID-19) Weekly Epidemiological Update – 31 August 2020. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200831-weekly-epi-update-3.pdf?sfvrsn=d7032a2a_4
2. E. Dong, H. Du, L. Gardner, *Lancet Infect. Dis.* **20**, 533–534 (2020).
3. D. Wrapp *et al.*, *Science* **367**, 1260–1263 (2020).
4. A. C. Walls *et al.*, *Cell* **181**, 281–292.e6 (2020).
5. M. Hoffmann, H. Kleine-Weber, S. Pöhlmann, *Mol. Cell* **78**, 779–784.e5 (2020).
6. M. Hoffmann *et al.*, *Cell* **181**, 271–280.e8 (2020).
7. J. Shang *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11727–11734 (2020).
8. T. Teesalu, K. N. Sugahara, V. R. Kotamraju, E. Ruoslahti, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16157–16162 (2009).
9. H. F. Guo, C. W. Vander Kooi, *J. Biol. Chem.* **290**, 29120–29126 (2015).
10. A. Plein, A. Fantin, C. Ruhrberg, *Microcirculation* **21**, 315–323 (2014).
11. R. Hollandi *et al.*, *Cell Syst.* **10**, 453–458.e6 (2020).
12. M. W. Parker, P. Xu, X. Li, C. W. Vander Kooi, *J. Biol. Chem.* **287**, 11082–11089 (2012).
13. B. J. Janssen *et al.*, *Nat. Struct. Mol. Biol.* **19**, 1293–1299 (2012).
14. A. Jarvis *et al.*, *J. Med. Chem.* **53**, 2215–2226 (2010).
15. J. K. Millet, G. R. Whittaker, *Virology* **517**, 3–8 (2018).
16. M. Simons, E. Gordon, L. Claesson-Welsh, *Nat. Rev. Mol. Cell Biol.* **17**, 611–625 (2016).
17. H. B. Pang *et al.*, *Nat. Commun.* **5**, 4904 (2014).
18. M. Ackermann *et al.*, *N. Engl. J. Med.* **383**, 120–128 (2020).
19. S.-Y. Lau *et al.*, *Emerg. Microbes Infect.* **9**, 837–842 (2020).

ACKNOWLEDGMENTS

We thank the Bristol Synthetic Biology Centre and the Advanced Computing Research Centre for provision of HPC (Bluegem), and the University of Bristol Wolfson Bioimaging Facility. We thank the University of Queensland Remote Operation Crystallisation and X-ray facility (UQ-ROCX) and the staff for their support with the crystallization experiments, and the staff of the Australian Synchrotron for assistance with x-ray diffraction data collection.
Funding: J.L.D. was supported by a Wellcome Trust studentship from the Dynamic Molecular Cell Biology Ph.D. program (203959/Z/16/Z), C.A.P. was supported by Beca Fundación Ramón Areces Estudios Postdoctorales en el Extranjero, and M.K.W. was supported by an MRC grant (MR/R020566/1) awarded to A.D.D. This project has received funding from the MRC (MR/P018807/1), Wellcome Trust (104568/Z/14/2), Lister Institute of Preventive Medicine, and Elizabeth Blackwell Institute for Health Research Rapid Response Call (COVID-19) awarded to P.J.C, the European Research Council under the European Union's Horizon 2020

research and innovation program (No 856581 - CHUbVi), and from MRC-AMED (MR/T028769/1) awarded to Y.Y., the Swiss National Science Foundation and Kanton Zurich awarded to U.F.G. B.M.C. is supported by an Australian National Health and Medical Research Council (NHMRC) Senior Research Fellowship (APP1136021) and Project Grant (APP1156493), and the United States Food and Drug Administration grant no. HHSF2232015101O4C "Ebola Virus Disease: correlates of protection, determinants of outcome and clinical management" amended to incorporate urgent COVID-19 studies awarded to J.A.H., A.D.D., and D.A.M. R.H. and P.H. acknowledge support from the LENDULET-BIOMAG Grant (2018-342), from H2020-discovAIR (874656), and from Chan Zuckerberg Initiative, Seed Networks for the HCA-DVP. T.T. was supported by the European Regional Development Fund (Project no. 2014-2020.4.01.15-0012), by European Research Council grant GLIOGUIDE and Estonian Research Council (grants PRG230 and EAG79, to T.T.). **Author contributions:** J.L.D., B.S., A.H., P.J.C., and Y.Y. conceived the study. J.L.D., B.S., K.K, and Y.Y. performed most of the experiments. K.K., M.K.W., D.A.M., and A.D.D. performed all work with infectious SARS-CoV-2 supervised by A.D.D. M.K.W. and A.D.D. isolated SARS-CoV-2 strains used for the work. K.C., C.A.P., M.B., L.S.G., U.F.G., K.K., R.B.S., D.K.S., J.A.H., and T.T. did experimental work and/or provided essential reagents. R.H. and P.H. performed image analysis. B.S., A.D.D., B.M.C., P.J.C., and Y.Y. supervised the research. J.L.D., B.S., A.D.D., P.J.C., and Y.Y. wrote the manuscript and made the figures. All authors read and approved the final manuscript. **Competing interests:** T.T. is an inventor of patents on CendR peptides and a shareholder of Cend Therapeutics Inc., a company that holds a license for the CendR peptides and is developing the peptides for cancer therapy. J.A.H. is a member of the Department of Health, New and Emerging Respiratory Virus Threats Advisory Group (NERVTAG) and the Department of Health, Testing Advisory Group. U.F.G. is a consultant to F. Hoffmann–La Roche Ltd, Switzerland. All other authors declare no competing interests. **Data and materials availability:** Coordinates and structure factors for the NRP1 b1-S1 CendR peptide complex have been deposited at the Protein Data Bank (PDB) with accession code 7JJC. All other data are available in the manuscript or the supplementary materials. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/. This license does not apply to figures/photos/artwork or other content included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/861/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S3
References (*20*–*33*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

14 June 2020; accepted 12 October 2020
Published online 20 October 2020
10.1126/science.abd3072

PREBIOTIC CHEMISTRY

Prebiotic synthesis of cysteine peptides that catalyze peptide ligation in neutral water

Callum S. Foden*, Saidul Islam*, Christian Fernández-García, Leonardo Maugeri, Tom D. Sheppard, Matthew W. Powner†

Peptide biosynthesis is performed by ribosomes and several other classes of enzymes, but a simple chemical synthesis may have created the first peptides at the origins of life. α-Aminonitriles—prebiotic α–amino acid precursors—are generally produced by Strecker reactions. However, cysteine's aminothiol is incompatible with nitriles. Consequently, cysteine nitrile is not stable, and cysteine has been proposed to be a product of evolution, not prebiotic chemistry. We now report a high-yielding, prebiotic synthesis of cysteine peptides. Our biomimetic pathway converts serine to cysteine by nitrile-activated dehydroalanine synthesis. We also demonstrate that *N*-acylcysteines catalyze peptide ligation, directly coupling kinetically stable—but energy-rich—α-amidonitriles to proteinogenic amines. This rare example of selective and efficient organocatalysis in water implicates cysteine as both catalyst and precursor in prebiotic peptide synthesis.

Peptides and proteins are essential to all life on Earth, and their biosynthesis is achieved by a highly evolved system of enzyme catalysts (*1*, *2*). Although the origin of protein synthesis in biology remains a mystery, life's exploitation of peptides may have predated the evolution of the complex enzymes that are now required to coordinate biosynthesis. Simple chemical processes could have furnished prebiotic peptides and peptide catalysts (*3*). To elucidate such chemical mechanisms, it is essential that we reflect on the biochemical strategies of extant biology to inform the systematic evaluation of prebiotic chemistry. For example, we recently demonstrated that α-peptide synthesis in water could be achieved by H_2S-mediated stoichiometric ligation (*4*). Our synthesis exploited a biomimetic N- to C-terminal chain-growth mechanism. This overcame long-standing problems that had prohibited the coupling of α-aminonitriles (*5*–*7*) and avoided the irrevocable side-chain modifications caused by electrophilic agents required to activate amino acids (*8*, *9*). Further reflection on the deep-seated role of thiols in nonribosomal peptide synthesis and core metabolism (*10*, *11*) has now led us to hypothesize that cysteine may have originated as a secondary product of sulfide-mediated peptide synthesis. We suspected that cysteine could be used to deliver catalytic peptide ligation (CPL) in water and a rare example of prebiotic synthesis directly generating catalytic function.

Cysteine is the primary organic source of sulfide in biology and the feedstock for essential cofactors such as glutathione and coenzyme A (CoA) (*12*). It is also an important residue within enzyme active sites, with vital functions in catalysis, redox sensing, and electron transfer, as well as being an essential ligand in ancient iron-sulfur proteins (*11*, *13*, *14*). It therefore seems almost inconceivable that cysteinyl thiols were not present during the development of nascent biological processes on early Earth, and yet this is not the prevailing view (*15*–*18*). Numerous, unsuccessful attempts to synthesize and isolate cysteines under prebiotically plausible conditions (*7*, *19*–*21*) have led to a widely held belief that cysteine is a biological invention (*17*, *18*), as well as a late addition to the genetic code (*15*–*18*). Here, we report a high-yielding prebiotic synthesis of cysteines and demonstrate that these cysteine peptides catalyze nonenzymatic CPL in neutral water. Our results support the hypothesis that cysteine was available at the origins of life as a secondary product of serine nitrile (Ser-CN) synthesis and that cysteines would have been a cornerstone of early catalytic activity.

The Strecker synthesis of aminonitriles is widely believed to play an important role in the prebiotic origins of amino acids, and we had previously identified glycolaldehyde (GCA) and β-mercaptoacetaldehyde (BMA) as key nodes in the chemical network required for abiogenesis of RNA and peptides (*3*, *22*, *23*), and specifically BMA as a Strecker precursor of cysteine (*22*). However, although Strecker reactions are generally highly efficient (*23*, *24*), BMA forms intractable and insoluble mixtures in Strecker reactions (fig. S1) (*7*). This is in stark contrast to GCA, which undergoes the Strecker reaction in excellent yield to afford a stable aminonitrile product, Ser-CN (fig. S2) (*23*, *24*). The observed disparity between congeners GCA and BMA is likely due to the rapid reaction of β-aminothiols with nitriles in water (*25*), and because cysteine nitrile (Cys-CN) is both a β-aminothiol and an α-aminonitrile, it is inherently unstable (*7*). These observations suggested to us that whereas the Strecker reaction of GCA could yield serine under prebiotic conditions, the Strecker reaction of BMA would not play a role in prebiotic cysteine synthesis. To overcome this inherent stability problem, we considered an alternative biomimetic pathway for cysteine synthesis. We suspected that clues to the prebiotic synthesis of cysteine might remain embedded within the extant biological pathway and that GCA, rather than BMA, was the key prebiotic precursor of cysteine.

The principal mechanism by which reduced inorganic sulfur is incorporated into bioorganic compounds is through cysteine biosynthesis (*17*, *18*). In plants and various archaea and bacterial species, cysteine biosynthesis begins with the enzymatic conversion of serine to *O*-acetylserine (Ser^{Ac}) [or *O*-phosphoserine (Sep)] followed by a pyridoxal-5′-phosphate (PLP)–dependent acetic acid β-elimination, followed by sulfide β-addition before disassociation of the cysteine-enzyme complex (Fig. 1A). Throughout this sequence, the α-amine remains covalently bound to PLP to promote elimination and to prevent the release and rapid decomposition of highly unstable dehydroalanine (Dha) (*26*). The nitrile equivalent of Dha, Dha nitrile (Dha-CN), has been synthesized by Eschenmoser and co-workers but was found to be extremely unstable and failed to react with H_2S to produce Cys-CN, instead undergoing rapid degradation, even under anhydrous conditions (*27*). However, we have recently shown that *N*-acylation of α-aminonitriles is a crucial element in initiating peptide synthesis in water, and it prevents hydantoin-, diketopiperazine-, and imidazole-induced peptide degradation (*4*–*6*). Similarly, *N*-acylation would stabilize Dha-CN. *N*-acylation would simultaneously prevent the highly favorable, but unwanted, enamine-to-imine tautomerization that precludes sulfur addition to Dha (*27*) and prevent the degradation of Cys-CN that is brought about by its free α-amine (*7*). Consequently, we recognized that serine diacylation presents a simple biomimetic strategy for prebiotic Dha synthesis, and we identified *N*,*O*-diacetyl-serine nitrile (Ac-Ser^{Ac}- CN) as a key intermediate for prebiotic cysteine synthesis (Fig. 1B). In addition to the electron-withdrawing effects of the α-nitrile, *N*,*O*-diacetylation would further enhance the acidity of the α-proton of Ac-Ser^{Ac}-CN as well as activate the serine hydroxyl moiety as a leaving group. We envisaged that these combined effects would promote Ac-Dha-CN synthesis at neutral pH, without recourse to the highly alkaline (pH > 13) conditions typically required for Dha formation in water (*28*) that would also promote peptide degradation (*29*).

We have a long-standing interest in (prebiotic) acylation (*30*, *31*), but it was not clear if the hydroxyl moiety of Ser-CN could be

Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, UK.
*These authors contributed equally to this work.
†Corresponding author. Email: matthew.powner@ucl.ac.uk

selectively acetylated in water. Therefore, we were pleased to observe chemoselective *N,O*-acetylation of Ser-CN with thioacetic acid (AcSH) and ferricyanide (*4*, *30*) to produce Ac-Ser^{Ac}-CN in up to 91% yield within 1 hour at room temperature (supplementary materials). Acetylation of Ser-CN with *N*-acetylimidazole (NAI) (*30*, *31*) was equally effective for Ac-Ser^{Ac}-CN synthesis in neutral water. We initially observed *O*-acetylserine nitrile (Ser^{Ac}-CN) as the major product (<10 min) with NAI (fig. S8) and attribute the nucleophilicity of the β-hydroxyl of Ser-CN to the pronounced electron-withdrawing effect of the α-nitrile and the low basicity (pK_{aH}) of this amino alcohol (supplementary materials). We also observed the formation of a stable Dha, Ac-Dha-CN, at near-neutral pH for the first time during these acetylation reactions. Optimal formation of Ac-Dha-CN from Ac-Ser^{Ac}-CN occurred at pH 8, and after 4 days at room temperature, an 85% conversion to Ac-Dha-CN was observed (Fig. 1B and fig. S30). This elimination is all the more notable because *N,O*-diacetylserinamide (Ac-Ser^{Ac}-NH$_2$) underwent near-exclusive hydrolysis at pH 8 (figs. S34 and S35) (*32*). The switch in reactivity between α-nitrile (acetate elimination) and α-amide (acetate hydrolysis) demonstrates the benefits of α-nitrile activation for acetate elimination and indicates that Ac-Dha-CN is predisposed to form in near-neutral water (*28*, *29*).

We next investigated Ac-Dha-CN synthesis via phosphate elimination. Acetylation of Sep-CN (*33*) in water with AcSH and ferricyanide at pH 7 yielded Ac-Sep-CN (80%) after 1 hour. Phosphoserines typically require alkaline pH and Ba^{2+} to promote Dha formation (*29*), but the formation of Ac-Dha-CN (8%) was nonetheless observed after heating Ac-Sep-CN at 60°C for 3 days at pH 7. The sluggish rate of phosphate elimination was enhanced by Mg^{2+}, yielding Ac-Dha-CN (24%) after 1 day at 60°C (fig. S33). This demonstrates that prebiotically plausible Dha formation can be achieved by serine acetylation or phosphorylation.

Ac-Dha-CN was found to be highly stable, and we did not observe the addition of acetate, phosphate, or hydroxide (even at pH 11) to this Dha, setting the stage for selective addition of inorganic sulfur to synthesize cysteine.

Ac-Dha-CN underwent near-quantitative conversion to Ac-Cys-SNH$_2$ upon incubation with H$_2$S, yielding the first prebiotic synthesis of a stable cysteinyl residue (Fig. 1B and fig. S36). We next investigated the prebiotic acetylating agent, AcSH, as a more water-soluble sulfide source at pH 7. Incubation of Ac-Dha-CN with AcSH led to quantitative thioester formation after 12 hours in phosphate buffer (Fig. 1B and fig. S38). The cysteine residue was then rapidly liberated by ammonolysis yielding Ac-Cys-CN (77%; Fig. 1B and fig. S44) or thiolysis to yield Ac-Cys-SNH$_2$ in 95% yield (Fig. 1B and fig. S39). Prolonged incubation of Ac-Cys^{Ac}-CN with H$_2$S gradually yielded cysteine thioacid Ac-Cys-SH and amide Ac-Cys-NH$_2$ (2:1) (figs. S47 and S48). High-yielding Ser-CN→Cys^{Ac}-CN conversion was also observed for *N*-acetylvalinylserine nitrile (Ac-Val-Ser-CN) (supplementary materials), demonstrating the efficacy of cysteine synthesis within a sterically encumbered peptide substrate.

Acetyl-CoA is biosynthesized from cysteine and is a universally conserved acetylating agent for protein, carbohydrate, and lipid metabolism. The relative simplicity of thioester Ac-Cys^{Ac}-CN suggests it may have been exploited as an acetyl-CoA analog and an activated source of acetate in (proto)metabolism (*10*), so we verified the proficiency of Ac-Cys^{Ac}-CN as a source of activated acetate. We suspected that acetyl transfer from Ac-Cys^{Ac}-CN to α-aminonitriles (AA-CN) would be promoted by their low pK_{aH} (e.g., Gly-CN pK_{aH} = 5.4), which renders AA-CN neutral and highly nucleophilic at neutral pH (*4*). To this end, we incubated Ac-Dha-CN, AcSH, and Gly-CN in phosphate buffer at pH 7 and room temperature. We observed in situ thioester formation and near-quantitative acetyl transfer from Ac-Cys^{Ac}-CN to Gly-CN, producing Ac-Gly-CN (81%) and Ac-Gly-SNH$_2$ (13%) after 3 days (fig. S67).

Prebiotic syntheses, much like biosyntheses, are necessarily multistep chemical pathways and, therefore, like all multistep processes, are susceptible to diminished overall yields unless the individual reactions proceed with high efficiency (*3*). In biosynthesis, enzyme catalysis achieves exquisite selectivity and high yields, but modern enzymes are a product of billions of years of evolution. In the absence of enzymes, prebiotic chemistry had to initially exploit unevolved and directly accessible alternatives, such as organocatalysts (*34*). Small-molecule catalysts could have functioned as rudimentary "enzymes" at the origins of life, and although they have been highly sought after since the watershed rediscovery of proline catalysis, progress has been greatly hampered by the generally poor activity of organocatalysts in water (*35*). However, the importance of nitriles at the origins of life (*3*), as well as their low background reactivity, suggested they may be ideal substrates for a highly selective and high-yielding CPL in water and warranted further investigation.

Peptide fragment ligations, such as native chemical ligation, are important reactions in

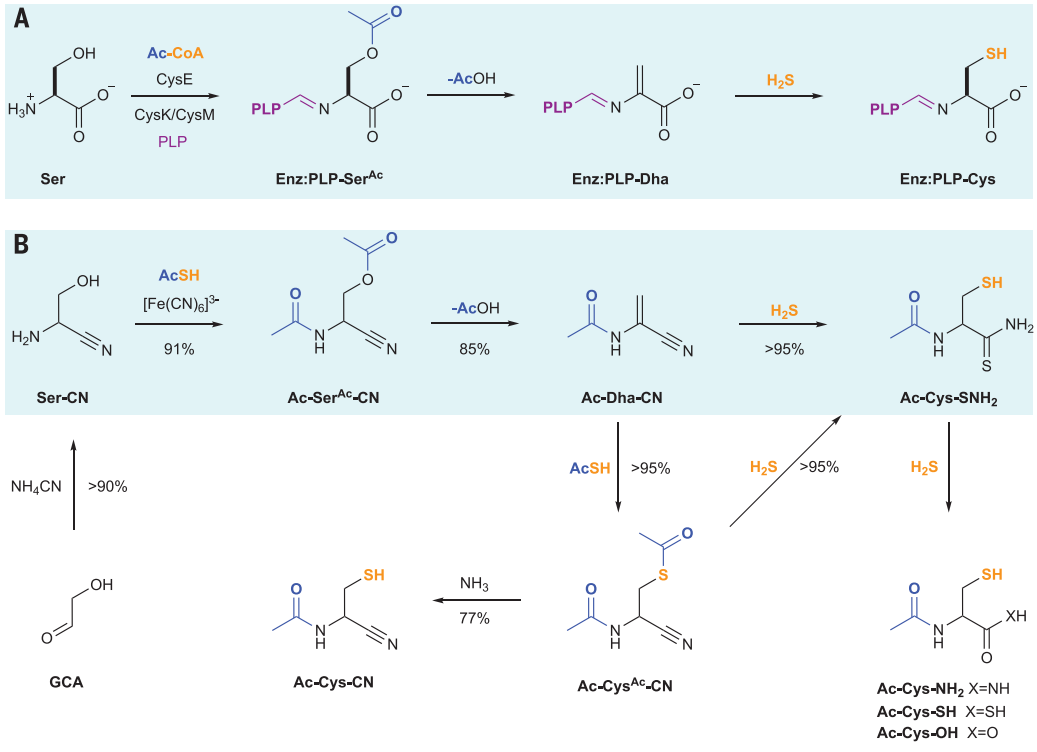


Fig. 1. Prebiotic cysteine synthesis. Biomimetic conversion of serinyl nitriles to cysteinyl nitriles. (**A**) Biosynthesis: PLP-dependent enzymatic cysteine synthesis pathway. (**B**) Prebiotic synthesis: Nitrile-activated synthesis of cysteines in water starting from Ser-CN, the stable Strecker product of GCA. Enz, enzyme; Ac, acetyl.

chemical biology and synthetic chemistry, facilitating rapid synthesis of longer peptides from smaller subunits (*36*–*38*). Fragment ligations have been proposed to play an important role in prebiotic chemistry (*39*), but the plausibility of these are diminished by the need for synthetically prepared C-terminal thioesters (*40*). We recently demonstrated high-yielding prebiotic fragment ligations (*4*) using stoichiometric H_2S and an activating agent (such as ferricyanide or cyanoacetylene) (Fig. 2A, i). Ideally, prebiotic fragment ligations through direct coupling of nitrile and amine fragments would sidestep thioamide and thioacid intermediates (*4*). Therefore, we became intrigued by the potential of thiols, such as *N*-acylcysteines, to act as organocatalysts for CPL. Reversible thiol addition to an α-amidonitrile could render peptide synthesis catalytic and redox-neutral via a highly reactive, but transient, thioimidate intermediate (Fig. 2A, ii). We anticipated that these thioimidates would be substantially more reactive than the thioesters conventionally exploited in peptide ligations (*36*, *37*) but would also be inherently protected from hydrolysis by reversible thiol-to-nitrile addition and elimination back to the stable nitrile substrate.

Eschenmoser and co-workers had previously reported a cysteine-catalyzed ammonolysis of an α-amidonitrile in methanol (*41*) to produce a C-terminal primary amidine. In water, however, the addition of cysteine to α-amidonitriles near-quantitatively yields thiazolines (supplementary materials), following the mechanism implicated in Cys-CN self-degradation (*7*). Thiazoline formation is, however, completely suppressed by cysteine *N*-acylation (*42*), suggesting that our biomimetic cysteine synthesis—which necessarily yields *N*-acyl-cysteines—is predisposed to furnish catalytically active cysteines. Therefore, we next tested Ac-Cys-OH as a catalyst for peptide ligation. We incubated Ac-Gly-CN and glycine with Ac-Cys-OH and observed CPL in water for the first time. This reaction yielded peptidyl amidine Ac-Gly^N-Gly-OH (60%) after 24 hours at 60°C and pH 7 (fig. S85). No activating agents were required to induce ligation. Moreover, peptide ligation no longer requires a C-terminal thioacid (*4*) or thioester and an N-terminal cysteine residue as the nucleophilic ligation partner (*36*, *37*, *40*). A broad spectrum of cysteine derivatives were excellent catalysts for peptidyl amidine synthesis in water, including Ac-Cys-NH_2 (69%) and *N*-acylcysteine peptides (57 to 71%), as well as CoA (65%) and simple thiols such as coenzyme M (79%) (fig. S85). Catalysis is essential to promoting this ligation, and the catalytic potency of simple thiols is highlighted by the lack of background reactivity. In control experiments (with no thiol catalyst), no coupling was observed, and Ac-Gly^N-Gly-OH (<1%) could only be detected after 7 days (fig. S86). Very little cysteine-catalyzed hydrolysis was observed in the absence of an amine nucleophile; incubating Ac-Gly-CN with Ac-Cys-OH (30 mol %, pH 7, 60°C) resulted in only 6% hydrolysis after 24 hours (fig. S87). These results underscore the outstanding kinetic stability of α-amidonitriles and the generality of thiol-catalysis, with only a severely hindered tertiary thiol failing to catalyze CPL (fig. S85).

CPL is specific and selective for proteinogenic α-peptide synthesis (Fig. 3). For example, the reaction of Ac-Gly-CN and Ac-β-Ala-CN (1:1; 200 mM) with glycine (200 mM) and Ac-Cys-OH (30 mol %) results in exclusive α-amidonitrile coupling to furnish Ac-Gly^N-Gly-OH (65%) with no detectable β-amidonitrile coupling of Ac-β-Ala-CN. We also observed only α-ligation upon challenging CPL with *N*-acetylglutamine dinitrile (Ac-Glx-CN). Finally, we only observed coupling of proteinogenic alanine in competition with α,α-disubstituted (nonproteinogenic amino acid) α–aminoisobutyric acid (Aib).

Thiol-catalyzed coupling of α-amidonitriles with α–amino acids is highly general; all proteinogenic α–amino acids coupled with Ac-Gly-CN to give peptidyl amidines Ac-Gly^N-AA-OH in good yields at pH 7 and 60°C (Table 1). However, dipeptides derived from serine, threonine, and asparagine underwent pronounced amidine hydrolysis to the corresponding peptides (Ac-Gly^N-AA-OH→Ac-Gly-AA-OH). In the reactions of serine and threonine, we observed oxazoline intermediates, suggesting that intramolecular catalysis by the amino acid side chain was responsible for rapid

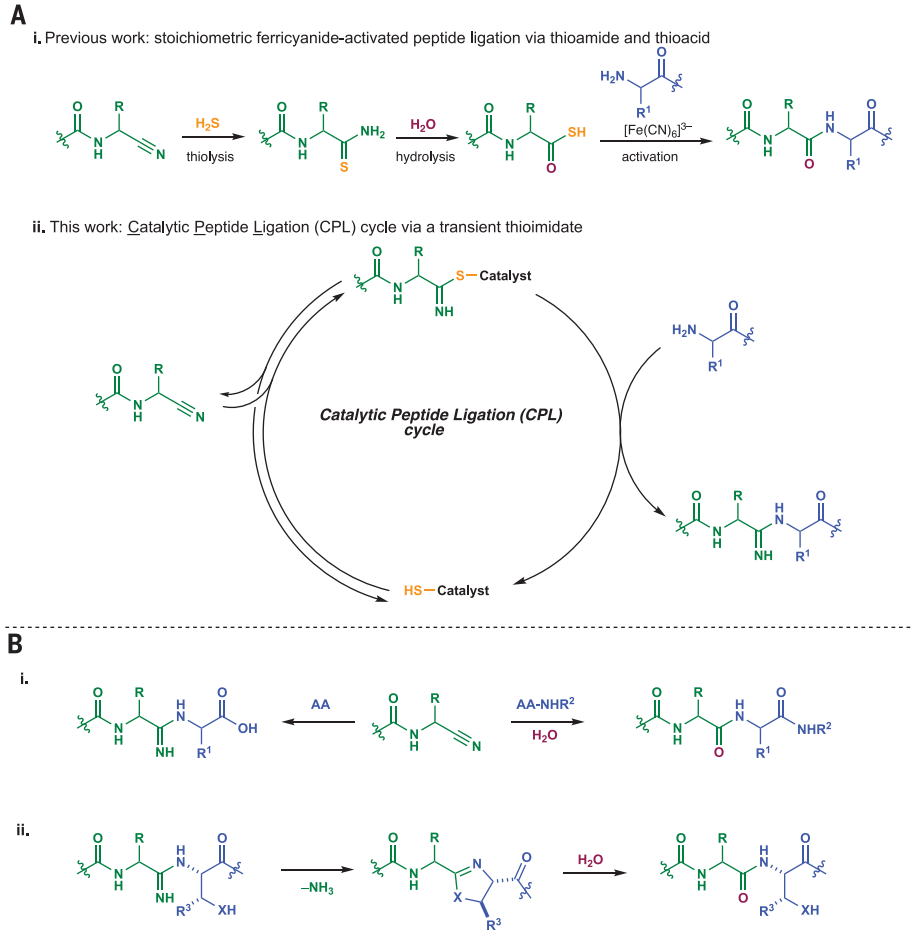


Fig. 2. Thiol-catalyzed peptide synthesis in neutral water. (**A**) (i) Previous work: Stoichiometric ferricyanide-activated peptide thioacid ligation, which requires temporally separated thiolysis and activation steps (*4*). (ii) This work: Thiol-catalyzed peptide ligation via a transient thioimidate intermediate that is chemoselectively intercepted by an amine nucleophile in water to form a peptidyl amidine directly from a stable nitrile without any activating agents. R, aminoacyl side chain; R_1, aminoacyl side chain. (**B**) (i) Peptidyl amidines persist if α-amidonitriles are coupled with an α–amino acid (AA), except serine, threonine, and cysteine (see Fig. 2B, ii). Intramolecular amide-assisted hydrolysis of the peptidyl amidine yields the native peptide bond (e.g., asparagine or R^2 = H or peptide). (ii) Intramolecular cyclization of serinyl (R^3 = H; X = O), threoninyl (R^3 = CH_3; X = O), or cysteinyl (R^3 = H; X = S) residues promote stereoretentive hydrolysis of peptidyl amidines to native peptides.

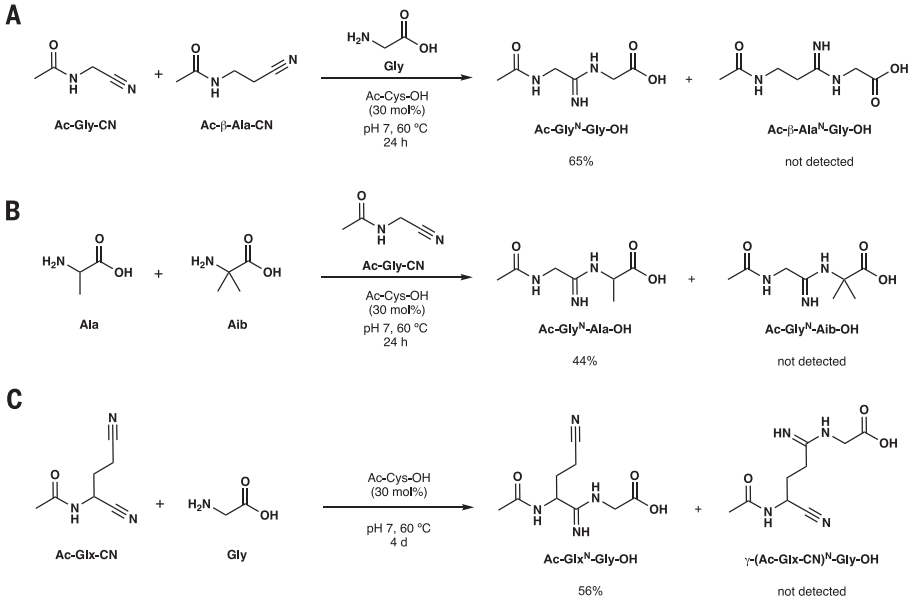


Fig. 3. Chemo- and regioselective organocatalytic proteinogenic α-peptide ligation. (**A** to **C**) CPL selectively yields native α-peptidyl bonds. Selective ligation of (A) Ac-Gly-CN (200 mM) with glycine (200 mM) affords Ac-Gly^N-Gly-OH in a stoichiometric competition with Ac-β-Ala-CN (200 mM), (B) Ac-Gly-CN (200 mM) with alanine (200 mM) affords Ac-Gly^N-Ala-OH in a stoichiometric competition with Aib (200 mM), and (C) Ac-Glx-CN (200 mM) with glycine (200 mM) affords Ac-Gly^N-Gly-OH.

Table 1. Organocatalytic ligation of *N*-acetylglycine nitrile with α–amino acids and α–amino amides. Yields for Ac-Cys-OH (30 mol %)–catalyzed formation of peptidyl amidines (Ac-Gly^N-AA^1-X) and peptides (Ac-Gly-AA^1-X) by coupling of Ac-Gly-CN (200 mM) with AA^1-X (1 equiv., pH 7, 60°C, 24 hours). See tables S6 and S7 for improved yields obtained by varying reaction conditions and further examples with other α-amidonitriles Ac-AA-CN. "–" indicates not observed.

AA^1-OH	Ac-Gly^N-AA^1-OH (%)	Ac-Gly-AA^1-OH (%)	AA^1-NH$_2$	Ac-Gly^N-AA^1-NH$_2$ (%)	Ac-Gly-AA^1-NH$_2$ (%)
Gly	60	2	Gly	21	52
DL-Ala	43	–	D-Ala	3	63
Arg	37	–	Arg	14	56
Asn	9	45	Asn	–	72
Asp	58	–	Asp	6	58
Gln	56	–	Gln	–	43
Glu	58	–	Glu	–	64
His	73	–	His	–	67
Ile	55	–	Ile	12	47
Leu	53	–	D-Leu	5	54
Lys	70*	–	Lys	25†	52‡
DL-Met	72	–	Met	5	62
Phe	21	–	Phe	8	52
Pro	58	–	Pro	–	21
Ser	–	61	Ser	–	68
Thr	–	51	Thr	–	69
Trp	32	5	Trp	4	45
Tyr	20	–	Tyr	3	62
Val	42	6	Val	7	50

*N^2-(Ac-Gly^N)-Lys-OH (43%), N^6-(Ac-Gly^N)-Lys-OH (24%), N^2,N^6-bis(Ac-Gly^N)-Lys-OH (3%). †Only the N^6-amidine, N^6-(Ac-Gly^N)-Lys-NH$_2$, was observed. ‡Only the N^2-peptide, N^2-(Ac-Gly)-Lys-NH$_2$, was observed.

amidine hydrolysis (Fig. 2B). Having also observed that the peptidyl amidine derived from asparagine hydrolyzed, we envisaged that amino amides (and, therefore, peptides) would behave similarly. We next examined whether amides intramolecularly catalyze amidine hydrolysis by coupling proteinogenic α–amino amides (AA-NH$_2$). All AA-NH$_2$ couplings resulted in selective dipeptide (Ac-Gly-AA-NH$_2$) synthesis, irrespective of their side chain (Table 1). This observed hydrolysis is important because, although peptidyl amidines undergo racemization as expected (*41*), N-terminal serine, threonine, and cysteine undergo stereoretentive coupling (Fig. 2B). The origins of biological homochirality remain a formidable challenge, requiring an as yet unknown symmetry-breaking event, and until that event, prebiotic syntheses must produce racemic mixtures (*43*). The observed racemization and epimerization of peptidyl amidines may have limited impact in a racemic (prebiotic) environment; however, the stereoretentive couplings of peptides with nucleophilic side chains and the intramolecular amide-catalyzed hydrolysis of peptidyl amidines offer a route to investigate dynamic kinetic resolution of peptide stereochemistry that may underpin peptide chiral resolution in oligomers, rather than α–amino acid (or α-aminonitrile) monomers (*44*).

Having demonstrated the tolerance of all proteinogenic aminoacyl residues at the ligation junction in thiol-catalyzed dipeptide synthesis, we explored the feasibility of organocatalytic peptide fragment ligations. We recently demonstrated the prebiotic synthesis of *N*-acylpeptide nitriles (Ac-AA$_n$-CN) in water by sulfide-mediated iterative ligation of α-aminonitriles (Fig. 2A, i) (*4*), and therefore

Table 2. Organocatalytic peptidyl-nitrile to peptide fragment ligations. Yields for MPA (160 mM)–catalyzed ligation of Ac-Gly$_3$-CN (100 mM) and peptides AA$_n$-OH (100 mM, 24 hours, 60°C, pH 7.0).

AA$_n$-OH	Ac-Gly$_3$-AA$_n$-OH (%)
Met-Gly	80
Ala-Ala-Ala	90
Ala-Gly-Ala	84
Gly-Ala-Gly	87
Gly-Gly-Gly	89
Gly-Gly-His	89
Leu-Leu-Leu	76*
Met-Ala-Ser	77
Phe-Gly-Gly	77†

*Forty-eight hours. †Ac-Gly$_2$-Gly^N-Phe-Gly$_2$-OH (12%) was also observed. Total ligation yield was 89%.

we chose to investigate Ac-Gly_3-CN as a prebiotically plausible peptide nitrile in fragment ligations. Peptide nitrile Ac-Gly_3-CN (100 mM) readily coupled with various peptides with 3-mercaptopropionic acid (MPA) as a catalyst at neutral pH to give excellent yields of peptides, even with stoichiometric (1:1) coupling partners (Table 2). The simplicity and generality of CPL underscore the privileged nature of thiol-catalyzed peptide ligations in water.

We have combined systems chemistry (*3*, *45*) with analysis of the strategy of modern biosynthesis to resolve two long-standing conundrums at the origins of life: the chemical origins of cysteine and catalytic peptide synthesis. We note that extant cysteine biosynthesis (*17*, *18*) bears a strong resemblance to the prebiotic synthesis we have outlined. However, the nonenzymatic reactions described here are contingent on an α-nitrile, rather than complex enzymes and cofactors. The α-amidonitrile moiety is not only prebiotically plausible but also ideally poised to activate Dha formation at near-neutral pH and provides the in-built energy required to drive peptide synthesis. The α-amidonitrile has a balance of kinetic stability and thermodynamic reactivity to deliver aqueous ligation without requiring any electrophilic activation, and the highly selective and reversible reaction that thiols undergo with α-amidonitriles makes thiol-catalyzed nitrile ligation a powerful and selective system for peptide ligation in water. The amidine intermediates observed during CPL suggest that ligation occurs by intermolecular addition of the nucleophilic coupling partner to a catalyst-bound thioimidate (Fig. 2A, ii). Our ligation [in contrast to thioester acylations (*36*, *37*)] tolerates all amino acid side-chain residues without protection to deliver high yields in neutral water and is not limited to ligation of N-terminal cysteine residues.

It is notable that a single amino acid residue, cysteine, provides robust catalysis for peptide ligation in water. The inherent catalytic activity of simple cysteinyl peptides makes them an excellent candidate for catalysts for protometabolic reactions in an abiotic environment. Finally, our data support a scenario in which nitriles served as an early energy currency on the primordial Earth, perhaps acting as a forerunner to adenosine 5′-triphosphate and thioesters that drive reactions in extant biology (*10*, *11*).

REFERENCES AND NOTES

1. P. Nissen, J. Hansen, N. Ban, P. B. Moore, T. A. Steitz, *Science* **289**, 920–930 (2000).
2. M. A. Fischbach, C. T. Walsh, *Chem. Rev.* **106**, 3468–3496 (2006).
3. S. Islam, M. W. Powner, *Chem* **2**, 470–501 (2017).
4. P. Canavelli, S. Islam, M. W. Powner, *Nature* **571**, 546–549 (2019).
5. M. S. Chadha, L. Replogle, J. Flores, C. Ponnamperuma, *Bioorg. Chem.* **1**, 269–274 (1971).
6. I. Shalayel, S. Coulibaly, K. D. Ly, A. Milet, Y. Vallée, *Life* **8**, 47 (2018).
7. I. Shalayel *et al*., *Eur. J. Org. Chem.* **2020**, 3019–3023 (2020).
8. L. Leman, L. Orgel, M. R. Ghadiri, *Science* **306**, 283–286 (2004).
9. H. Griesser, M. Bechthold, P. Tremmel, E. Kervio, C. Richert, *Angew. Chem. Int. Ed.* **56**, 1224–1228 (2017).
10. C. De Duve, *Blueprint for a Cell: The Nature and Origin of Life* (N. Patterson, Burlington, NC, 1991).
11. J. E. Goldford, H. Hartman, T. F. Smith, D. Segrè, *Cell* **168**, 1126–1134.e9 (2017).
12. L. B. Poole, *Free Radic. Biol. Med.* **80**, 148–157 (2015).
13. R. V. Eck, M. O. Dayhoff, *Science* **152**, 363–366 (1966).
14. C. Bonfio *et al*., *Nat. Chem.* **9**, 1229–1234 (2017).
15. J. T.-F. Wong, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1909–1912 (1975).
16. E. N. Trifonov, *Gene* **261**, 139–151 (2000).
17. A. Sauerwald *et al*., *Science* **307**, 1969–1972 (2005).
18. K. Fujishima *et al*., *Sci. Rep.* **8**, 1776 (2018).
19. B. N. Khare, C. Sagan, *Nature* **232**, 577–579 (1971).
20. A. L. Weber, S. L. Miller, *J. Mol. Evol.* **17**, 273–284 (1981).
21. E. T. Parker *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **108**, 5526–5531 (2011).
22. M. W. Powner, S.-L. Zheng, J. W. Szostak, *J. Am. Chem. Soc.* **134**, 13889–13895 (2012).
23. S. Islam, D.-K. Bučar, M. W. Powner, *Nat. Chem.* **9**, 584–589 (2017).
24. B. H. Patel, C. Percivalle, D. J. Ritson, C. D. Duffy, J. D. Sutherland, *Nat. Chem.* **7**, 301–307 (2015).
25. Z. Chen *et al*., *Angew. Chem. Int. Ed.* **59**, 3272–3279 (2020).
26. A. J. Borchert, D. C. Ernst, D. M. Downs, *Trends Biochem. Sci.* **44**, 849–860 (2019).
27. G. Ksander *et al*., *Helv. Chim. Acta* **70**, 1115–1172 (1987).
28. I. Photaki, *J. Am. Chem. Soc.* **85**, 1123–1126 (1963).
29. M. F. Byford, *Biochem. J.* **280**, 261–265 (1991).
30. F. R. Bowler *et al*., *Nat. Chem.* **5**, 383–389 (2013).
31. C. Fernandez-Garcia, M. W. Powner, *Synlett* **28**, 78–83 (2017).
32. B. M. Anderson, E. H. Cordes, W. P. Jencks, *J. Biol. Chem.* **236**, 455–463 (1961).
33. A. J. Coggins, M. W. Powner, *Nat. Chem.* **9**, 310–317 (2017).
34. C. F. Barbas III, *Angew. Chem. Int. Ed.* **47**, 42–47 (2008).
35. M. P. van der Helm, B. Klemm, R. Eelkema, *Nat. Rev. Chem.* **3**, 491–508 (2019).
36. P. E. Dawson, T. W. Muir, I. Clark-Lewis, S. B. Kent, *Science* **266**, 776–779 (1994).
37. S. S. Kulkarni, J. Sayers, B. Premdjee, R. J. Payne, *Nat. Rev. Chem.* **2**, 0122 (2018).
38. V. R. Pattabiraman, J. W. Bode, *Nature* **480**, 471–479 (2011).
39. J. Söding, A. N. Lupas, *BioEssays* **25**, 837–846 (2003).
40. D. H. Lee, J. R. Granja, J. A. Martinez, K. Severin, M. R. Ghadiri, *Nature* **382**, 525–528 (1996).
41. H. Moser *et al*., *Helv. Chim. Acta* **69**, 1224–1262 (1986).
42. U. E. W. Lange, B. Schäfer, D. Baucke, E. Buschmann, H. Mack, *Tetrahedron Lett.* **40**, 7067–7070 (1999).
43. J. E. Hein, D. G. Blackmond, *Acc. Chem. Res.* **45**, 2045–2054 (2012).
44. D. Beaufils, G. Danger, L. Boiteau, J.-C. Rossi, R. Pascal, *Chem. Commun.* **50**, 3100–3102 (2014).
45. J. W. Szostak, *Nature* **459**, 171–172 (2009).

ACKNOWLEDGMENTS

We thank K. Karu, M. Puchnarewicz (UCL Mass Spectrometry Facility), E. Samuel (Mass Spectrometry, UCL School of Pharmacy), A. E. Aliev (NMR spectroscopy), P. Canavelli, B. Thoma, and J. Fairchild for preliminary work and experimental assistance. **Funding:** We thank the Engineering and Physical Sciences Research Council (EPSRC) (EP/K004980/1, EP/M507970/1, and EP/P020410/1), the Simons Foundation (318881FY19 and 493895), and the Volkswagen Foundation (94743) for financial support. **Author contributions:** M.W.P. conceived the research. C.S.F., S.I., C.F.-G., T.D.S., and M.W.P. designed and analyzed the experiments. C.S.F., S.I., C.F.-G., and L.M. conducted the experiments. S.I. assembled and wrote the supplementary materials with contributions from C.S.F. and C.F.-G. C.S.F. and S.I. contributed equally to this work and are listed alphabetically. S.I. and M.W.P. wrote the manuscript with assistance from T.D.S. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability:** All data are available in the main text and the supplementary materials.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/370/6518/865/suppl/DC1
Materials and Methods
Figs. S1 to S209
Tables S1 to S14
References (*46*–*51*)

29 June 2020; accepted 1 October 2020
10.1126/science.abd5680

Cell-Freezing Media

AMS Biotechnology announces two chemically defined, cryopreservation excipient solutions: STEM-CELLBANKER and HSC-BANKER. Available in both dimethyl sulfoxide (DMSO)-containing and DMSO-free formulations, STEM-CELLBANKER is a chemically defined freezing media optimized for stem-cell and induced pluripotent stem-cell storage as well as fragile primary cell storage. Published data supports its ability to cryopreserve organoids and tissues to allow the recovery of viable cells. Free from animal-derived components, this popular medium contains only chemically defined U.S. Pharmacopeia (USP)-, European Pharmacopeia (EP)-, and Japanese Pharmacopeia (JP)-grade ingredients. Manufactured to be completely free of serum and animal-derived components, HSC-BANKER contains only EP- or USP-grade ingredients, making it highly suitable for storage of hematopoietic stem cells developed for cell-therapy applications. Recently the master files of HSC-BANKER were accepted by the Center for Biologics Evaluation and Research within the U.S. Food and Drug Administration. HSC-BANKER is supplied ready-to-use and requires no special devices, such as a controlled-rate freezer, to achieve consistently high viabilities following resuscitation from cryopreservation, even over extended long-term storage. HSC-BANKER significantly increases cell viability while maintaining cell pluripotency, normal karyotype, and proliferation ability after freeze-thaw. It is evaluated for endotoxins, pH, osmolarity, and mycoplasma contaminants to ensure GMP equivalent quality.

AMS Biotechnology
For info: 617-945-5033
www.amsbio.com/cellbanker-cell-freezing-media

2D Field-Flow Fractionation

The EAF2000 series from Postnova Analytics is a powerful and field-proven 2D field-flow fractionation (FFF) system, which can be applied as either traditional asymmetric flow FFF or alternatively as a hybrid FFF system utilizing both separation forces in the same channel simultaneously. This allows particle size and molar mass separations induced by the crossflow field as well as charge separations by electrophoretic mobility. As particle and molecule charge play a primary role in many applications, such as protein aggregation, polymer flocculation, particle agglomeration, and pharmaceutical formulations, the EAF2000 enables significantly better understanding of these phenomena and will help to establish more efficient product development and QC processes.

Postnova Analytics
For info: 801-521-2004
www.postnova.com/product/systems/eaf2000-electrical-af4.html

Chromatography Flowmeter

TESTA Analytical Solutions has developed a new flowmeter that enables continuous measurement of flow rate without interference in chromatography systems. Flow rate is one of the most important parameters in any liquid chromatography system; it determines retention time or volume and has a major influence on reproducibility. Compatible with all high-performance liquid chromatography (HPLC) and gel permeation chromatography/size-exclusion chromatography (GPC/SEC) solvents, the TESTA flowmeter is conveniently sized and powers itself from a USB connection. A PC-based app allows continuous recording and storage of the measured flow rates. The current flow rate is also displayed on the flowmeter's integral high-resolution organic LED display, allowing easy control of current flow value. Extraordinary high resolution and wide dynamic range make the TESTA flowmeter the perfect flow-monitoring tool for the most demanding HPLC and GPC/SEC systems.

TESTA Analytical Solutions
For info: +49-(0)-30-864-24076
www.testa-analytical.com

Application-Optimized Machine Vision Lenses

Resolve Optics offers a service to design, develop, and supply application-optimized machine vision lenses for any industrial and robotic image-processing application. High-quality construction, coupled with precision engineering, results in outstanding machine vision lenses that deliver sharp, high-resolution, optically precise images. Our lenses provide wide fields of view with little or no distortion, optical designs and coatings that are balanced to give best performance at a desired wavelength or waveband, and compact lens designs where the target application is space-limited.

Resolve Optics
For info: +44-(0)-1494-777100
www.resolveoptics.com/machine-vision

In Vitro Pyrogen Testing

Lonza Bioscience announces the release of the PyroCell Monocyte Activation Test (MAT) System, a sustainable, reliable solution for in vitro pyrogen testing. The new offering ensures the safety of parenteral pharmaceuticals during development, manufacture, and product release. It provides sensitive pyrogen detection without the use of experimental animals, thereby supporting sustainability objectives while helping to deliver safe products to the market. Through a collaboration with Sanquin Reagents B.V., the PyroCell MAT System consists of pooled, cryopreserved peripheral blood mononuclear cells (PBMCs) specifically developed for use with the MAT. These PyroCell Kit PBMCs eliminate the need to qualify blood donors and undertake cell isolation for each test run. With the PyroCell MAT System, the cells can be available whenever the need arises.

Lonza
For info: +41-(0)-61-316-81-11
www.lonza.com/mat

CCD Cameras

The Atik VS series of scientific cooled charge-coupled device (CCD) cameras is flexible enough to meet the needs of both scientific OEMs and microscopy users. It provides a versatile solution for any low-light application, which simplifies the designer's task of integration and has an optional five-position integrated filter wheel allowing filters to be loaded without opening the camera. The VS series delivers the highest-quality images under low-light conditions, thanks to 16-bit digitization that offers 65,536 gray levels, plus high-performance thermoelectric cooling to –35°C below ambient. At 6 MP/s, the VS series offers the fastest digitization of any Atik camera. This range supports a wide variety of Sony CCD sensors, including ICX285, ICX274, ICX674, ICX655, ICX694, and ICX814. The VS255 and VS825 cameras are specifically for use with quantitative PCR instruments for accurate, reliable COVID-19 testing.

Atik Cameras
For info: 888-733-3034
www.atik-cameras.com/product/atik-vs-range

Electronically submit your new product description or product literature information! Go to www.sciencemag.org/about/new-products-section for more information.

Newly offered instrumentation, apparatus, and laboratory materials of interest to researchers in all disciplines in academic, industrial, and governmental organizations are featured in this space. Emphasis is given to purpose, chief characteristics, and availability of products and materials. Endorsement by *Science* or AAAS of any products or materials mentioned is not implied. Additional information may be obtained from the manufacturer or supplier.

The National Academies of | SCIENCES ENGINEERING MEDICINE

POSTDOCTORAL RESEARCH FELLOWS

Air Force Science & Technology Fellowship Program

The National Academies of Sciences, Engineering, and Medicine administers **postdoctoral and senior fellowship awards** at the U.S. Air Force Research Laboratory (AFRL), the U.S. Air Force Institute of Technology (AFIT), and the U.S. Air Force Academy (USAFA) under the Air Force Science & Technology Fellowship Program (AF STFP).

We are seeking highly qualified candidates who are U.S. citizens and hold, or anticipate earning, a doctorate in a variety of fields of science or engineering.

Application deadline dates (four annual review cycles):

- February 1
- May 1
- August 1
- November 1

Awardees have the opportunity to:

- Conduct independent research in an area compatible with the interests of the Air Force laboratories
- Devote full-time effort to research and publication
- Access the excellent and often unique Air Force research facilities
- Collaborate with leading scientists and engineers

Awardee benefits:

- Base stipend starting at $76,542; may be higher based on experience
- Health insurance (including dental and vision), relocation benefits, and professional travel allowance

For detailed program information, to search for AFRL, AFIT, and USAFA Research Opportunities, and to contact prospective Research Adviser(s), visit **www.nas.edu/afstfp**.

The National Academies of | SCIENCES ENGINEERING MEDICINE

POSTDOCTORAL RESEARCH ASSOCIATES

NRC Research Associateship Programs

The National Academies of Sciences, Engineering, and Medicine administers **postdoctoral and senior research awards** at participating federal laboratories and affiliated institutions at locations throughout the U.S and abroad.

We are seeking highly qualified candidates who hold, or anticipate earning, a doctorate in a variety of fields of science or engineering. Degrees from foreign universities should be equivalent in training and research experience to a doctoral degree from a U.S. institution. Citizenship eligibility varies among the sponsoring laboratories.

Application deadline dates (four annual review cycles):

- February 1
- May 1
- August 1
- November 1

Awardees have the opportunity to:

- Conduct independent research in an area compatible with the interests of the sponsoring laboratory
- Devote full-time effort to research and publication
- Access the excellent and often unique facilities of the federal research enterprise
- Collaborate with leading scientists and engineers at the sponsoring laboratories

Awardee benefits include:

- Stipends ranging from $45,000 to $86,000; may be higher based on experience
- Health insurance (including dental and vision), relocation benefits, and a professional travel allowance

For detailed program information, to search Research Opportunities, and to contact prospective Research Adviser(s) visit **www.nas.edu/rap**.

Science 2020
TOP EMPLOYER
Who's the top employer for 2020?
Science Careers' annual survey reveals the top companies in biotech & pharma voted on by Science readers.
Read the article and employer profiles at sciencecareers.org/topemployers

ScienceCareers online @sciencecareers.org

ScienceCcareers
FROM THE JOURNAL SCIENCE AAAS

Confused about your next career move?

Download Free Career Advice Booklets!

ScienceCareers.org/booklets

YOU'D SMILE TOO IF YOU JUST ADVANCED YOUR CAREER.

Find your next job at ScienceCareers.org

There's scientific proof that when you're happy with what you do, you're better at what you do. Access career opportunities, see who's hiring and take advantage of our proprietary career-search tools. Get tailored job alerts, post your resume and manage your applications all in one place: sciencecareers.org

ScienceCareers

FROM THE JOURNAL SCIENCE AAAS

By **Sasha McDowell**

Difficult conversations

It was 11 o'clock on a Friday morning—time to meet with my supervisor. Normally, I would have hurried over to my computer, coffee in hand, casually adjusting my bed head in order to appear at least somewhat professional. But this May morning was different. I tried to pull myself together, but I hadn't slept much, and nothing could relieve my puffy eyes. A few days earlier, George Floyd had been killed by the police. His homicide made me even more aware that I am a Black student at a predominantly white institution. To my knowledge, I was the only Black graduate student in my department at that time. Would I be able to talk about the killing with my white male supervisor? Did I want to talk about it with him? Or would we ignore the elephant in the room?

Prior to the COVID-19 outbreak, our meetings had typically been limited to research. When the time came, I would anxiously put together whatever data I had and head over to his office. Between grading assignments for the class in which I was a teaching assistant (TA), taking a course of my own, and learning difficult techniques in the lab, I didn't have much research progress to share, and I worried he thought I was unproductive. Yet I hesitated to tell him I felt overwhelmed for fear that I would seem incompetent.

"It was as though an invisible barrier had been broken."

After our lab closed in response to the pandemic, that dynamic shifted dramatically. Without new data to discuss, our meetings veered toward casual conversations about topics such as the latest news on the pandemic, how my supervisor was juggling teaching his young daughter from home while working, and whether my furniture from Ikea had finally arrived. It was as though an invisible barrier had been broken. He was no longer just the person overseeing my research; he was another human being trying to navigate the difficult times the COVID-19 pandemic had imposed on us all.

When we did discuss research, being separated by a screen somehow gave me the confidence to bring up challenging topics and ask questions that felt naïve. I was working on my Ph.D. proposal and wasn't yet adept at literature searches—something a supervisor might assume graduate students know how to do. Before the pandemic, I likely wouldn't have brought up my fears about overlooking crucial papers. But now I felt comfortable sharing my concerns and asking for advice. We also discussed how I could reserve time for research while also teaching, whether by finding a less time-intensive course to TA or making the most of the semesters when I wasn't teaching. It was refreshingly open discourse.

Despite this new casual camaraderie, the prospect of discussing George Floyd's killing was daunting. We had spoken about many global issues, but nothing as sensitive as this one. I calmed myself with the thought that it probably wouldn't come up.

But on that morning, my supervisor started our meeting a little differently from usual. Typically, he asked, "How's it going?" to which I'd routinely reply, "It's going OK." This time, though, he asked, "How are you *doing*?"

At first, it was awkward tiptoeing around race, but his genuine interest in my well-being made me feel as though I was in a safe space to discuss Black issues and Black representation in science. We acknowledged my minority status. I volunteered that I was lucky to not experience racism at our institution, but that I was interested in promoting equity, diversity, and inclusion (EDI) practices at our university. After this conversation, my supervisor seemed to make it his mission to connect me with others with similar goals. As a result, I am now part of a departmental trainee group to discuss such issues and a member of an EDI committee—which my supervisor decided to join as well.

Although the pandemic has restricted my research, it has unexpectedly strengthened my relationship with my supervisor. It has enabled me to become more involved in issues of social justice and, hopefully, to make a contribution to science and academia that goes beyond the lab bench. Just as important, I think the impact has been mutual. ■

Sasha McDowell is a graduate student at the University of British Columbia, Vancouver. Send your career story to SciCareerEditor@aaas.org.

ILLUSTRATION: ROBERT NEUBECKER

Published by AAAS